Pädagogisch-psychologische Diagnostik 1

Pädagogisch-psychologische Diagnostik

Band 1

Theoretische und methodische Grundlagen

von

Lothar Tent
und Ingeborg Stelzl

Hogrefe · Verlag für Psychologie
Göttingen . Bern · Toronto · Seattle

Prof. Dr. Lothar Tent, geb. 1928. Lehramtsstudium 1948-1952, Lehrer 1952-1960. Diplom-Psychologe 1958, Promotion 1962. 1962-1968 Wissenschaftlicher Assistent an der Universität Marburg. 1968 Habilitation im Fach Psychologie. 1968/69 Professur für Pädagogische Psychologie an der Universität Gießen. 1969 Professur für Sonderpädagogik an der Universität Marburg. Seit 1973 Professor am Fachbereich Psychologie, Universität Marburg.

Prof. Dr. Ingeborg Stelzl, geb. 1944. Studium der Psychologie, Philosophie und Physik; 1967 Promotion in Graz. Seit 1973 Professorin am Fachbereich Psychologie, Universität Marburg.

Druck- und Bindearbeiten: Dieterichsche Universitätsbuchdruckerei
W. Fr. Kaestner GmbH & Co. KG, D-3400 Göttingen-Rosdorf
Printed in Germany
Auf säurefreiem Papier gedruckt
ISBN 3-8017-0647-8

Vorwort

Der vorliegende Band enthält den ersten Teil einer zweibändigen Einführung in die Pädagogisch-psychologische Diagnostik. Diese Einführung wendet sich primär an Studierende mit Hauptfach Psychologie oder Erziehungswissenschaft im zweiten Studienabschnitt und ist als vorlesungsbegleitende Lektüre, aber auch als Leitfaden zur Prüfungsvorbereitung gedacht, insbesondere für die entsprechenden Ausschnitte der Fächer "Pädagogische Psychologie" und "Diagnostik" der Diplomprüfung Psychologie. Im vorliegenden Band I geht es um allgemeine theoretische und methodische Grundlagen, in Band II sollen Verfahren zu einzelnen inhaltlichen Bereichen vorgestellt und Anwendungsfelder im Sinn typischer diagnostischer Fragestellungen erörtert werden.

Zu den theoretischen und methodischen Grundlagen von Diagnostik gehören u.E.

(1) eine begriffliche Klärung dessen, was Diagnostik ist und soll,
(2) die testtheoretischen Grundlagen, auf deren Basis diagnostische Verfahren entwickelt und beurteilt werden, und schließlich
(3) eine Gegenstandsbestimmung und theoretische Einordnung dessen, was diagnostiziert werden soll, sowie eine Reflexion der praktischen Rahmenbedingungen und der rechtlichen Grundlagen für die Anwendung.

Dementsprechend gliedert sich der vorliegende Band in drei Teile. Die Teile I und III wurden von L. Tent verfaßt, Teil II wurde von I. Stelzl beigesteuert. Die vorgetragenen Positionen und Argumente werden insgesamt von beiden Autoren vertreten. Teil I schafft die ***begrifflichen Voraussetzungen*** und diskutiert die allgemeinen Grundlagen Pädagogisch-psychologischer Diagnostik: Geklärt werden soll, was unter Pädagogisch-psychologischer Diagnostik zu verstehen ist, wozu sie dient, wie sie vorgeht, was wir von ihr erwarten. Dazu wird in den Abschnitten 1.1 und 1.2 zunächst das Gebiet abgegrenzt und die Bedeutung von Diagnostik für das pädagogische Handeln herausgestellt. In 1.3 werden dann die bereits in der Alltagsdiagnostik enthaltenen allgemeinen Grundprinzipien von Diagnostik sichtbar gemacht und in 1.4 die wesentlichen Elemente, Annahmen und Probleme professioneller Diagnostik herausgearbeitet. Dabei spielt der Begriff des Merkmals eine zentrale Rolle. Es wird eine Systematik von Merkmalsklassen vorgestellt, weiter wird die Beziehung von Verhaltensmerkmalen zu Konstrukten und die Bedeutung von Konstrukten für die diagnostische Praxis behandelt. Verhalten wird dabei als Resultante aus Personmerkmalen und situativen Umweltbedingungen aufgefaßt, und die Varianz als zumindest im Prinzip diesen Varianzquellen entsprechend aufteilbar gedacht. Neben der Präzisierung der Merkmale werden die Präzisierung der Meßoperation (Standardisierung, Ökonomisierung, Meßgenauigkeit) und die Verifizierung diagnostischer Aussagen als wesentliche Kriterien professioneller Diagnostik herausgestellt, was dann in Teil II unter testtheoretischen Gesichtspunkten näher erörtert wird. Teil I schließt als Hauptergebnis mit einer zusammenfassenden Definition von Pädagogisch-psychologischer Diagnostik und einer Definition des Begriffs "psychologischer Test".

Teil II ist den ***testtheoretischen Grundlugen*** gewidmet. Ein Ziel dieses Abschnitts liegt darin, die Begriffe und Methoden zu erläutern, die in fast allen Testhandanweisungen auftauchen: Das sind zum einen Begriffe aus der klassischen Testtheorie, zum

anderen klassische multivariate Verfahren. Darüber hinaus soll ein Überblick über testtheoretische Entwicklungen und Kontroversen gegeben werden, soweit sie für die Pädagogisch-psychologische Diagnostik relevant sind.

Wenn in den Kapiteln 2 bis 5 klassische Testtheorie und multivariate Verfahren behandelt werden, so kann das hier sicher nicht in dem Umfang und auf dem mathematischen Niveau erfolgen, das man in der Diagnostikprüfung im Hauptdiplom Psychologie anstrebt. Dazu muß auf gesonderte Lehrveranstaltungen bzw. Lehrbücher verwiesen werden. Kapitel 2 ist als ein Leitfaden der wichtigsten Begriffe der klassischen Testtheorie zu lesen; die Kapitel über multivariate Verfahren sind auf elementarem Niveau gehalten, so daß sie auch von Studierenden ohne einschlägige Vorkenntnisse und Interessenten aus benachbarten Gebieten verstanden werden sollten. Ziel ist es, die Grundgedanken dieser multivariaten Verfahren und mögliche Anwendungen in der Pädagogisch-psychologischen Diagnostik so weit erkennbar zu machen, daß eine kritische Auseinandersetzung möglich ist. Trotz des einführenden Niveaus sollte der Text auch Lesern mit stärkeren Vorkenntnissen noch etwas zu bieten haben: Es wird Wert darauf gelegt, speziell die Punkte darzustellen, die Gegenstand von Kontroversen waren oder sind (z.B. die Populationsabhängigkeit der klassischen Gütekriterien oder die Rolle der Normalverteilung in Kapitel 2; die Einwände gegen die klassische Faktorenanalyse und die Grenzen der konfirmatorischen Faktorenanalyse im Kapitel 4), sowie auf Anwendungsgesichtspunkte einzugehen, die in rein formal orientierten Darstellungen manchmal zu kurz kommen (z.B. Kapitel 3: Fragestellungen und Fehlerrisiken bei der Interpretation von Differenzen in Testprofilen; Probleme bei der Interpretation von Gruppenprofilen als Anforderungsprofile; Kapitel 5: Können die Anforderungen an die klassischen Gütekriterien niedriger angesetzt werden, wenn der Test “nur” Forschungszwecken dient?).

Kapitel 6 enthält drei unterschiedliche Abschnitte, die mit der Absicht geschrieben sind, über Entwicklungen zu informieren, die nicht unbedingt zum Standardwissen aus einer Testtheorie-Veranstaltung gehören. Alle drei Themen bewegen sich im begrifflichen Ansatz der klassischen Testtheorie und haben jeweils einen spezifischen Bezug zur Pädagogisch-psychologischen Diagnostik: Die Theorie der Generalisierbarkeit (Abschnitt 6.1) bietet sich als begrifflicher Rahmen an, wenn z.B. bei Schulleistungstests repräsentative Aufgabenstichproben gezogen werden sollen, um dann zu fragen, wie von den vorliegenden Aufgaben auf die Grundgesamtheit aller Aufgaben generalisiert werden kann. In 6.2 wird die Diskussion um kriterienorientierte vs. normorientierte Messung dargestellt, die speziell im Zusammenhang mit der Konstruktion lehrzielorientierter Tests geführt wurde. In 6.3 geht es um den Versuch, mit methodischen Mitteln den Begriff der Testfairness (Fairness gegenüber sozial benachteiligten Gruppen) zu definieren und damit die Grundlage für empirische Untersuchungen zur Frage der Testfairness zu schaffen.

Kapitel 7 enthält die Grundzüge und wichtigsten Modelle der sog. probabilistischen Testtheorie (Latent-Trait-Modelle). Die Beschränkung der Darstellung fiel hier nicht leicht, weil gerade dieses Gebiet den Reiz der Aktualität hat und die Entwicklung noch nicht voll abgeschlossen ist. In Hinblick auf einen mäßigen Gesamtumfang und ein mittleres Anforderungsniveau erschien eine Behandlung von Detailfragen aus laufender Forschung nicht angebracht. Die Auswahl erfolgte primär unter dem Gesichtspunkt, daß Anwendungen aus dem Bereich der Pädagogisch-psychologischen Diagnostik bereits vorliegen oder sich unmittelbar abzeichnen sollten. Auf weitere Varianten und auf Verbindungen zwischen den Modellen wird nur hingewiesen.

Kapitel 8 behandelt adaptives Testen und baut insofern auf Kapitel 7 auf, als mit Hilfe der Latent-Trait-Modelle das Problem gelöst werden kann, wie Punktwerte trotz von Proband zu Proband unterschiedlicher Aufgabenauswahl zu vergleichen sind.

Das letzte Kapitel in Teil II ist Fragen der Veränderungsmessung gewidmet. In 9.1 werden zunächst formale Ansätze dargestellt, u.a. spezielle Modelle, die im Rahmen des Latent-Trait-Ansatzes zur Erfassung von Lernprozessen entwickelt wurden. Anschließend werden mehr inhaltlich orientierte Ansätze, wie die Vorschläge zur Konstruktion änderungssensitiver Tests und die Entwicklung spezieller Lerntests behandelt. Dabei wird in groben Zügen auch über inhaltliche Erfahrungen berichtet. In 9.2 wird die Evaluationsforschung als Anwendungsbereich besonders herausgegriffen. Hier tritt zu den testtheoretischen Fragen der Veränderungsmessung als Kernfrage das Problem hinzu, ob die diagnostizierten Veränderungen der zu evaluierenden Maßnahme zuzuschreiben sind. Anhand von drei Beispielen soll deutlich gemacht werden, welche methodischen Probleme dabei auftreten können und weshalb wissenschaftlich fundierte Evaluationsforschung nicht durch Alltagserfahrung und daran angelehnte "natürliche" Methoden ersetzt werden kann.

Nachdem in Teil I die allgemeinen begrifflichen und theoretischen Voraussetzungen und in Teil II die testtheoretischen Grundlagen behandelt wurden, wird in Teil III der Gegenstand Pädagogisch-psychologischer Diagnostik näher beleuchtet, und es werden allgemeine Probleme und Voraussetzungen der Anwendung diagnostischer Verfahren erörtert. Kapitel 10 behandelt den Begriff der Schulleistung als zentrales Konstrukt Pädagogisch-psychologischer Diagnostik und erläutert seine Beziehungen zu kognitiven, motorischen, sozialen, affektiven und motivationalen Lehrzielen. Es wird ein Bedingungsmodell für das Zustandekommen von Schulleistung vorgestellt, das als theoretische Grundlage und Interpretationsbasis dienen soll. Weiter werden allgemeine praktische Fragen angesprochen: die Frage nach dem zweckmäßigen Zeitpunkt und der Häufigkeit, mit der Diagnostik eingesetzt werden soll, Fragen nach Rückwirkungen diagnostischer Erhebungen auf den Schüler und auf den Lernprozeß, die Frage nach Fehlerquellen und Fehlschlüssen im Urteilsprozeß, denen sowohl der Lehrer in der alltäglichen Schülerbeurteilung als auch der Psychologe bei der Gutachtenerstellung unterliegen kann.

Kapitel 11 schließt diesen Band ab. Es behandelt berufsethische und rechtliche Aspekte. Es werden sowohl allgemeine Grundsätze als auch spezifische Rechtsvorschriften für Lehrer und Psychologen, sowie Fragen der rechtlichen Zuständigkeit erläutert.

Der geplante zweite Band wird sich mit einzelnen inhaltlichen Anwendungsbereichen befassen. Es sollen diagnostische Verfahren für bestimmte Bereiche besprochen werden: zur Diagnose kognitiver Lernvoraussetzungen (Schuleingangstests, allgemeine und spezielle Intelligenztests u.a.), zur Diagnose emotionaler, motivationaler und sozialer Lernvoraussetzungen (Persönlichkeits-, Einstellungs- und Interessentests), zur Diagnose von Wissen, Kenntnissen und Fertigkeiten (Schulleistungstests) sowie zur Diagnose spezieller Verhaltensauffälligkeiten. Weiter sollen die wichtigsten Anwendungsfelder im Sinn typischer diagnostischer Fragestellungen (Schuleingangsdiagnostik, Diagnose der Eignung für weiterführende Schulen, Sonderschulbedürftigkeit, Studieneignung und Hochschulzulassung, außerschulische Erziehungsberatung) behandelt werden. Schließlich soll zu Testkritik und Einwänden gegen Pädagogisch-psychologische Diagnostik Stellung genommen werden.

Wir haben folgenden Mitarbeiterinnen, die an der Fertigstellung von Band I beteiligt waren, zu danken: Frau Weskamm für das Anfertigen der Abbildungen und die Mithilfe beim Korrekturlesen, Frau Groll und Frau Schmitt für das Schreiben des Manuskripts.

Marburg, im' November 1992 Lothar Tent Ingeborg Stelzl

Inhaltsverzeichnis

1. Grundlegende Annahmen und Definitionen

1. **Welche Bedeutung hat die Diagnostik für pädagogisches Handeln, und wie läßt sich dies begründen?**
2. **Wodurch unterscheidet sich professionelle Diagnostik von Alltagsdiagnostik?**
3. **Auf welche grundlegenden Annahmen stützt sich Diagnostik, und wie gelangt man zu möglichst genauen und zutreffenden Aussagen?**
4. **Wie ist professionelle Pädagogisch-psychologische Diagnostik zweckmäßig zu definieren?**

Vorstrukturierende Lesehilfe

Erziehung und Unterricht sind Lebensbereiche, in denen Diagnostik eine besonders große Rolle spielt. Ständig müssen pädagogische Entscheidungen unterschiedlicher Tragweite getroffen werden. Ihre Wirksamkeit hängt u.a. davon ab, wie zutreffend die individuellen Lernvoraussetzungen und Fähigkeiten, aber auch die emotionale Verfassung und die motivationale Bereitschaft eines Schülers erkannt und berücksichtigt werden. Obwohl Differenzierung und Individualisierung seit langem als Prinzipien der Unterrichtsorganisation anerkannt sind, ist der Status der professionellen Pädagogisch-psychologischen Diagnostik hierzulande unbefriedigend.

Die wissenschaftlich fundierte Diagnostik, wie sie sich seit grob einhundert Jahren entwickelt hat, fußt auf der Alltagsdiagnostik. Die Alltagsdiagnostik ist jedoch in vieler Hinsicht unzulänglich. Die persönlichkeitstheoretischen und methodischen Konzepte, die tragfähige Lösungen für die meisten diagnostischen Fragestellungen möglich machen, werden erläutert. Eine wesentliche Rolle spielen dabei die Präzisierung der Merkmale und der Meßoperationen sowie die Verifizierung der diagnostischen Aussagen. Die instrumentehe Qualität der diagnostischen Hilfsmittel muß umso höheren Ansprüchen genügen, je mehr von den Entscheidungen abhängt, zu deren Begründung sie beitragen sollen. Das wichtigste Kriterium ist daher die ***empirische Validität*** der Methoden.

Abschließend wird Diagnostik als ein systematisches Vorgehen zur Gewinnung und Analyse von Merkmalsunterschieden an Personen definiert.

1.1 Bezeichnung des Gebiets

Die Bezeichnung ***"Pädagogisch-psychologisch"*** im Text dieses Buches bringt zum Ausdruck, daß die Diagnostik, um die es hier geht, beides zugleich ist, pädagogische und psychologische Diagnostik. Sie ist ***pädagogisch,*** weil ihre Fragestellungen aus der Erziehungspraxis stammen und weil unser Text sich auf diese Praxis bezieht. Überall da, wo es aus pädagogischen Gründen notwendig oder ratsam ist, die indivi-

duellen Bedingungen und die Ergebnisse menschlichen Lernens zu kennen, und immer wenn individuelles Verhalten mit pädagogischen Mitteln beeinflußt werden soll, ist Diagnostik unerläßlich. Diagnostische Tätigkeiten haben nicht nur einen beträchtlichen Anteil am Unterrichtsalltag, ihnen kommt auch ein hoher, kaum zu überschätzender Stellenwert zu: Ohne eine fundierte diagnostische Routine wäre ein professioneller, am Individuum orientierter Unterricht nicht möglich. Besonders sorgfältige Diagnosen sind vonnöten, wenn unerwartete Schwierigkeiten im Verhalten und in den Leistungen von Schülern auftreten und wenn wichtige, langfristig wirksame Entscheidungen über die pädagogische Behandlung von Schülern zu treffen sind, z.B. bei der Einschulung, bei Umschulungen, bei Kurszuweisungen oder auch beim "Sitzenbleiben".

Psychologisch ist diese Diagnostik deshalb, weil sie in der Regel Verhaltens- und Leistungsaspekte betrifft, die Gegenstand der theoretischen wie der empirischen Erkenntnisgewinnung in der Psychologie sind. Fachhistorisch kommt hinzu, daß die diagnostischen Methoden und Theorien, über die wir heute verfügen, vornehmlich von Psychologen entwickelt worden sind. Doch wäre es müßig, hier pädagogische und psychologische Anteile auseinanderzudividieren. Die moderne Pädagogisch-psychologische Diagnostik hat ihre Wurzeln selbstverständlich in der überkommenen pädagogischen Praxis. Aus deren diagnostischen Bedürfnissen ist sie entstanden, und sie dient nichts anderem, als eben diese Praxis zu verbessern.

Es liegt deshalb nahe, statt von Pädagogisch-psychologischer einfach von Pädagogischer Diagnostik zu sprechen. Diese verkürzende Bezeichnung ist vor etwa 20 Jahren von Ingenkamp (vgl. 1985, S. 10) in Analogie zum angloamerikanischen educational measurement/assessment eingeführt und von anderen Autoren aufgegriffen worden (z.B. Klauer, 1978; Süllwold, 1983). Die Verbindungen "heilpädagogische" bzw. "sonderpädagogische" Diagnostik waren ohnehin schon früher geläufig. Gegen die praktische Kurzform "Pädagogische Diagnostik" ist nichts einzuwenden, solange man sich bewußt bleibt, daß Pädagogische Diagnostik in der Regel - mit jeweils unterschiedlicher Akzentuierung - auch psychologische Diagnostik ist.

Im Unterschied zur ärztlichen Diagnostik, die vorwiegend auf die Feststellung (oder den Ausschluß) krankhafter Abweichungen vom "Normal''-Zustand des Gesunden oder Unauffalligen gerichtet ist, umfaßt die Pädagogische Diagnostik die gesamte Spannbreite der vorkommenden Zustände und Zusammenhänge, die für die individuelle pädagogische Behandlung ***aller*** Schüler bedeutsam sind. Zwar werden die erhobenen Daten auch in der Pädagogischen Diagnostik auf Vergleichswerte bezogen, doch ziehen nicht nur die auffälligen, von einer Norm abweichenden, sondern grundsätzlich ***alle*** Werte je unterschiedliche pädagogische Konsequenzen nach sich. Von wenigen Ausnahmen abgesehen, ist in der Pädagogik - anders als in der Medizin - stets eine "Behandlung", nämlich eine bestimmte pädagogische Maßnahme oder "Intervention", angezeigt.

In den folgenden Abschnitten werden einige allgemeine pädagogische und psychologische Grundlagen der Diagnostik kurz dargestellt.

1.2 Allgemeine pädagogische Grundlagen

Ganz gleich, wer in dieser Praxis tätig ist, jede diagnostische Aktivität ist eingebunden in ein konkretes pädagogisches Handlungsfeld. Dem Pädagogen erscheint dies

selbstverständlich; der Psychologe, der nicht zugleich auch pädagogisch ausgebildet und erfahren ist, muß sich erst darauf einstellen und entsprechende Kenntnisse erwerben. In seiner allgemeinsten Bedeutung wird pädagogisches Handeln zumeist als "Erziehen" bezeichnet. In diesem weiteren Sinne ist Erziehung als das ***intentionale Herbeiführen relativ dauerhafter Veränderungen von Personmerkmalen durch mentale Beeinflussung*** zu verstehen. ***Veränderung von Personmerkmalen*** bedeutet, individuelle Ist-Zustände in neue, vorgegebenen Soll-Werten entsprechende oder angenäherte Ist-Zustände zu überführen. ***Mentale Beeinflussung*** heißt, daß die intendierten Veränderungen über kognitive Prozesse der Informationsaufnahme und -Verarbeitung herbeigeführt werden, die ihrerseits vom emotionalen und motivationalen Zustand des Lernenden abhängen. Ziel der Erziehung ist es, mengentheoretisch veranschaulicht, die Schnittmenge der Verhaltens- und Erlebenselemente von Lehrenden und Lernenden systematisch zu vergrößern. Pädagogische Effekte gehen aber nicht nur von der gezielten interpersonalen Beeinflussung aus. Andere Umweltbedingungen, namentlich soziakulturelle Faktoren, wie Vorbild- und Medienwirkungen, tragen ebenfalls, wenn auch eher beiläufig, zum Erwerb oder zur Veränderung von Merkmalen bei (sog. ***funktionale*** Erziehung).

Diese knappe Definition genügt für unsere Zwecke. Im Unterschied zu manchen pädagogischen Definitionen von Erziehung ist sie insofern wertungsfrei, als sie über Wert oder Unwert der Ziele, Intentionen und Ergebnisse des pädagogischen Handelns keine Vorentscheidung trifft. Sie ist zugleich offen in Richtung auf pädagogisch-psychologische ***Verhaltensmodifikation*** oder ***Therapie,*** von der wir sprechen, wenn es um die mentale Beeinflussung von Ausgangszuständen mit Krankheitswert geht. Im Regelfall stellt Erziehen eine Art ***asymmetrischer sozialer Interaktion*** dar, deren beabsichtigte (oder unbeabsichtigte) Wirkung sich stets auf Individuen richtet und nur an Individuen manifest werden kann.

Der pädagogische Grundsatz, dabei die individuelle Eigenart der Kinder zu beachten, ist alt. Ideengeschichtlich geht er in der Neuzeit auf den künstlerischen und intellektuellen Individualismus der Renaissance und des Humanismus zurück. Zahlreiche Denker haben seither die pädagogische Bedeutung des ***Individualitätsprinzips*** hervorgehoben und präzisiert. Schon der große Pädagoge Arnos Comenius (1592-1670), dem es eigentlich um eine breit angelegte Massenerziehung ging, hat gefordert, neben dem Alter auch die Unterschiede in der Begabung (Veranlagung) und im Lernfortschritt der Kinder systematisch zu berücksichtigen. Er formuliert bereits didaktisch ***differenzierte*** Vorschriften, die im wesentlichen bis heute Bestand haben. Dem Kulturkritiker und Aufklärer Jean Jacques Rousseau (1712-1778) wird die psychologisch bedachtsame Förderung der natürlichen Individualität des Kindes zum Angelpunkt für eine tiefgreifende Umgestaltung der Gesellschaft. Der bekannte Schweizer Erzieher Johann Heinrich Pestalozzi (1746-1827) hat dann den zentralen pädagogischen Begriff der "Individuallage" geprägt. Darin werden die zeitgeschichtlichen, die gesellschaftlichen und materiellen Umweltverhältnisse, in die ein Kind hineingeboren wird, mit den personalen Aspekten seiner Konstitution, seiner Urteilsfähigkeit, seiner Motivation und seiner Charaktereigentümlichkeiten zu einem dynamischen Gesamtkonzept vereint. Dessen psychologischer Kern besteht aus der Art und Weise, wie sich das Kind in seiner sozialen Umwelt erlebt. Pestalozzi fordert, daß alle Erziehung von der ***Individuallage*** des Kindes auszugehen hat. In bewußter Abkehr von jeglichem Schematismus und Formalismus in der Pädagogik wurde schließ-

lich der Ruf "Vom Kinde aus" zur sinnfälligsten Maxime der Schulreformbewegung des ausgehenden 19. und beginnenden 20. Jahrhunderts (s. Dietrich, 1982).

Mit der zunehmenden Verwissenschaftlichung vieler Lebensbereiche werden in dieser Zeit auch die ersten diagnostischen Verfahren im heutigen Sinne entwickelt. Mit ihrer Hilfe sollte die alte Forderung nach der Differenzierung des Unterrichts besser in die Praxis umgesetzt werden, als dies aufgrund des üblichen Rückgriffs auf die Alltagserfahrung von Lehrern und Eltern möglich erschien. Bezeichnenderweise vollzog sich ein wesentlicher Teil dieser frühen Entwicklung innerhalb einer neuen Forschungsrichtung, die von ihren Verfechtern als "experimentelle Pädagogik" bezeichnet wurde (Ernst Meumann, 1862-1915; Wilhelm August Lay, 1862-1926). Sie zielte darauf ab, die Pädagogik insgesamt zu einer (möglichst) exakten Wissenschaft auszubauen und die Erziehungspraxis auf ein empirisch gesichertes Fundament zu stellen (vgl. z.B. Lay, 1903; Meumann, 1907. Beide gaben seit 1905 die ***Zeitschrift für Experimentelle Pädagogik*** heraus).

Differenzierung und Individualisierung sind heute unumstrittene Organisationsprinzipien für Schule und Unterricht. Die Pädagogisch-psychologische Diagnostik hat seit der Jahrhundertwende, vor allem in den USA, einen bemerkenswerten Aufschwung genommen. Trotz bedeutender Anteile deutscher Autoren an dieser Entwicklung und trotz erheblicher Fortschritte in den letzten Jahrzehnten ist die Akzeptanz der standardisierten diagnostischen Verfahren (Tests und Fragebogen) im deutschen Sprachraum hinter den Möglichkeiten, die sie bieten, aus verschiedenen Gründen zurückgeblieben (Tent, 1969, S. 28-33; Ingenkamp, 1985, S. 257-264). Zu diesen Gründen zählen gewisse Unzulänglichkeiten, die selbst bei sorgfältiger Konstruktion der Verfahren in Kauf genommen werden müssen, sowie ideologische Vorbehalte gegenüber der Funktion solcher Hilfsmittel und Probleme mit der sachgerechten Verwertung ihrer Ergebnisse (zur Kritik an der Pädagogischen Testdiagnostik vgl. Ingenkamp, 1989). Bis jetzt ist die wissenschaftlich begründete Diagnostik bei uns weder in der Unterrichtspraxis noch in der Lehrerausbildung zur Selbstverständlichkeit geworden.

Der Verwissenschaftlichung von Praxis, allzumal der pädagogischen, sind sicher Grenzen gesetzt. Aber auch in der Pädagogik erfordert professionelles Handeln, praktische Entscheidungen wo immer möglich auf empirisch gesicherte Erkenntnisse zu stützen. Wenn man hier, wie andernorts, das Prinzip der Optimierung von Handlungsentscheidungen anerkennt, gibt es keine rationalen Gründe, auf dafür geeignete Erkenntnismittel zu verzichten. Die Unvollkommenheit der Instrumente ist kein Gegenargument, solange nachweislich bessere Alternativen nicht zur Verfügung stehen und die bekannten Eigenschaften und Schwächen der Verfahren angemessen beachtet werden. Der allgemeine triviale Grundsatz, das jeweils Bestmögliche zu tun, gilt hier ebenso uneingeschränkt wie der Grundsatz, empirisch und nicht ideologisch zu entscheiden, was unter gegebenen Bedingungen und im Hinblick auf gegebene Ziele das Bestmögliche ist. Dazu muß man die in Betracht kommenden Alternativen gründlich genug kennen.

1.3 Alltagsdiagnostik

Diagnostische Urteile zu bilden, ist ein alltäglicher Vorgang. Wir beobachten und registrieren die Erscheinung, das Auftreten und die sprachlichen Äußerungen anderer,

schätzen das Aufgenommene ein und schließen daraus auch auf "die Person". Wir trachten danach, uns ein Bild vom anderen zu machen. Wir möchten wissen, mit wem wir es zu tun haben, "was für ein Mensch" der andere ist, was wir künftig von ihm zu erwarten haben und wie wir uns am besten darauf einstellen. Wir haben erkannt, daß unser eigenes Handeln auf ganz bestimmte Fähigkeiten, Zielvorstellungen, Bedürfnisse und Interessen zurückgeht, und zumeist können wir die (oft widerstreitenden) Beweggründe angeben, die uns veranlaßt haben, dieses oder jenes zu tun oder zu lassen. Entsprechende Gedanken machen wir uns darüber, weshalb sich andere so und nicht anders verhalten. Wir versuchen u.U., uns in den anderen "hineinzuversetzen". Wir wollen ihn so gut wie möglich verstehen. Denn Verständnis verschafft uns Sicherheit; es vermittelt den Eindruck, dem anderen gerecht zu werden, besser mit ihm auszukommen und gezielter auf ihn einwirken zu können. Wir vergleichen zu diesem Zweck die Menschen miteinander und orientieren uns dabei an Maßstäben, die unserer eigenen Erfahrung oder dem überkommenen "diagnostischen Regelwissen" entstammen (z.B. "Stille Wasser gründen tief", "Der Apfel fallt nicht weit vom Stamm", "An seinem Umgang erkennt man den Menschen" usw.).

Diese Vorgänge werden uns am ehesten dann bewußt, wenn es sich um jemanden handelt, mit dem wir in der Folgezeit häufig zusammentreffen wollen oder müssen, etwa wenn wir uns Freunde oder Partner aussuchen oder wenn z.B. ein Lehrer eine neue Klasse übernimmt, bzw. Schüler einen neuen Lehrer bekommen. Wir neigen dazu, wiederholte Beobachtungen zu verallgemeinem, und je besser wir einen Menschen zu kennen glauben, desto mehr gehen wir dazu über, ihm ganz bestimmte Fähigkeiten und Eigenschaften zuzuschreiben. Die Alltagssprache enthält eine nahezu unerschöpfliche Fülle von Ausdrücken und Wortkombinationen, mit deren Hilfe wir Menschen, uns selbst eingeschlossen, "charakterisieren" können. Jemand, bei dem uns dies nicht recht gelingt, bleibt uns fremd; wir bezeichnen ihn dann z.B. als "verschlossen", "in sich gekehrt" oder "undurchsichtig".

Selbstkritische Beobachter stellen in der Tat fest, daß es oft nicht einfach ist, zu schlüssigen Ergebnissen zu gelangen. Sie machen die Erfahrung, daß nicht nur einzelne Menschen, sondern auch einzelne Verhaltensbereiche unterschiedlich gut einzuschätzen sind. Sie wissen, daß sie sich täuschen können und sind deshalb auch nicht überrascht, wenn das Erwartete nicht eintrifft. Sie hüten sich, vorschnell zu urteilen und zu verallgemeinem. Sie bleiben bei vorsichtiger Vermutung, wo der Eilfertige sich schon eine feste Überzeugung bildet. Sie bedenken, daß Menschen, vor allem Kinder, sich mitunter stark und manchmal schnell verändern können. Sie differenzieren von Fall zu Fall zwischen Verhaltensweisen, bei denen sich Änderungen oder Schwankungen bemerkbar machen, und solchen, die vergleichsweise stabil erscheinen.

Dem aufmerksamen Beobachter entgeht auch nicht, daß sich Menschen vielfach unterschiedlich verhalten, je nachdem, wem sie gegenüberstehen, in welcher Situation und in welcher Verfassung sie sich befinden. Der einzelne Beurteiler ist kaum in der Lage, die ganze Spannbreite des Verhaltens anderer zu überblicken. Es kommt hinzu, daß die Fähigkeit, das Verhalten unserer Mitmenschen "richtig" zu deuten, offenbar nicht jedem in gleicher Weise gegeben ist.

Zwar entwickeln wir schon als Kinder, im ständigen Austausch mit den Menschen in unserer Umgebung, ein diagnostisches "Gespür", und mit wachsender Erfahrung verfeinern sich unsere diagnostischen Fertigkeiten und werden reichhaltiger. Und doch dauert es offenbar lange, bis wir jemanden "wirklich" kennen. So jedenfalls

belehrt uns das Sprichwort vom Scheffel Salz, den man erst mit ihm gegessen haben muß. Aber können wir uns überhaupt auf unsere Menschenkenntnis verlassen? Ist nicht jeder Mensch so unverwechselbar und einmalig mit seinem individuellen Werdegang und seiner vielschichtigen Umwelt, daß jeder Versuch, ihm diagnostisch gerecht zu werden, grundsätzlich scheitern muß? Die Literatur ist voller Widersprüche. Sie spricht vom ewigen Rätsel, das der Mensch dem Menschen bleibe, aber auch davon, daß das Erfassen nur eines Menschen in seiner ganzen Tiefe den Weg dazu eröffne, alle zu verstehen (s. Kasten S. 21).

1.4 Professionelle psychologische Diagnostik

Der Rückgriff auf die Alltagspraxis soll verdeutlichen, daß die wissenschaftlich begründete Diagnostik von heute in der Tat aus dieser Praxis hervorgegangen ist und nach wie vor enge Beziehungen dazu hat. In den Anfängen ihrer gut hundertjährigen Geschichte stehen konkrete Fragestellungen aus dem klinischen und pädagogischen, aus dem betrieblichen, dem forensischen und dem militärischen Bereich im Vordergrund. Doch spielen auch mehr theoretische Interessen an interindividuellen Unterschieden schon frühzeitig eine Rolle. Ihre theoretischen Grundlagen werden teils in unmittelbarem Zusammenhang mit der diagnostischen Praxis, teils erst nachträglich entwickelt.

Alle wesentlichen Elemente, Annahmen und Probleme der professionellen Diagnostik sind in der diagnostischen Alltagspraxis vorgebildet oder lassen sich daraus herleiten. Beide stützen sich auf das in den Grundzügen gleiche Persönlichkeitsmodell. Danach kommt allen Menschen-ungeachtet ihrer Individualität-dasselbe Spektrum psychophysiologischer Funktionen, Erlebnisqualitäten und Verhaltenskategorien zu, und danach lassen sich im beobachtbaren Verhalten Zusammenhänge erkennen, die man in Regeln fassen kann. Beide gehen auf dasselbe menschliche Grundbedürfnis nach Überschaubarkeit und Sicherheit der Lebensvollzüge zurück, und beide dienen demselben allgemeinen Zweck: menschliches Verhalten “kalkulierbar” zu machen und durch möglichst zutreffende Vorhersagen die Ungewißheit im zwischenmenschlichen Umgang zu reduzieren.

Diagnostik ist selbstverständlich nur ein Teil unseres Bemühens, Ungewißheit zu verringem. In erster Linie wird dies durch überindividuell verbindliche, einheitliche Verhaltensrichtlinien wie Sitten und Gebräuche, Verträge, Regeln, Normen und Sanktionen bewirkt. Von daher wäre “individuelles” Verhalten eher als ein nicht weiter reduzierbarer Unsicherheitsfaktor zu verstehen, der sich der Prognostizierbarkeit grundsätzlich entzieht. Für die partielle Gültigkeit dieses Arguments scheint die Tatsache zu sprechen, daß die psychologische Diagnostik zwar die relativ besseren Informationen liefert, in weiten Bereichen aber nach wie vor keine befriedigende Vorhersage machen kann.

Daraus zu folgern, man solle in der psychologischen und pädagogischen Praxis auf gründliche Diagnosen verzichten, wäre ein Fehlschluß. Die allgemeine Erfahrung und die vielfach belegte Gültigkeit diagnostischer Methoden sprechen gegen eine solche Auffassung. In wichtigen Teilbereichen menschlichen Verhaltens können wir offenbar mit hinreichender intraindividueller Kontinuität und mit hinreichender Beständigkeit interindividueller Differenzen rechnen. Hinreichend heißt, daß eine darauf ge-

Wie gut ist die Alltagsdiagnostik?

(1) Meili (in Meili & Steingrüber, 1978, S. 25-26) hat die *Intelligenz* dreier sechsjähriger Jungen mit deutlich verschiedenen IQ unter drei Bedingungen schätzen lassen. Die Beurteiler konnten sich auf Standfotos oder Filmaufnahmen stützen oder das reale Verhalten der Jungen beobachten. Sie sollten die Jungen hinsichtlich ihrer Intelligenz in eine Rangordnung bringen. Die Ergebnisse waren nicht besser als der Zufall. Die drei Darbietungsarten unterschieden sich darin nicht (s. Tabelle).

Beurteilung der Intelligenz nach Foto, Film und Natur; Einstufung in % (Richtige Beurteilungen kursiv).

Intelligenz	Darbietungsart: Foto gut	Foto mittel	Foto schwach	Film gut	Film mittel	Film schwach	Natur gut	Natur mittel	Natur schwach
Bester (IQ 115)	28	62		17	50	33	12		41
Mittlerer (IQ 100)		14	21	62	25	29	46	47	17	36
Schwacher (IQ 77)		58	17	25	58	21	21	41	36	23

(2) Schwieriger ist es, die Brauchbarkeit der Beurteilung von *Temperamentsmerkmalen* (Persönlichkeitseigenschaften i.e.S.) zu überprüfen. Cohen (1969, Abschn. II) ließ mit Hilfe von 15 bipolaren Schätzskalen, die "Dominanz", "Beliebtheit" und "Gewissenhaftigkeit" erfaßten, sowohl persönlich bekannte als auch unbekannte Personen beurteilen. Von den unbekannten Personen lagen Fotos, Handschriften und Selbsteinschätzungen vor. Bei Einschätzung aufgrund persönlicher Bekanntschaft stimmten die Beurteiler zwar mit r = 0.34 im Mittel signifikant höher überein als bei Einschätzung anhand von Fotos, Handschrift und Selbsteinstufung; dies reicht jedoch als Grundlage für die Gültigkeit von Eindrucksurteilen nicht aus. Auch bei simultaner Beurteilung mehrerer Informationsquellen schwankten die Korrelationen im Mittel lediglich zwischen r = 0.03 und maximal 0.27.

(3) Die Ergebnisse der Forschung zur Zuverlässigkeit und Gültigkeit (Validität) diagnostischer Eindrucksurteile hat Merz (1963, S. 44) wie folgt zusammengefaßt:

"1. Die *Zuverlässigkeit von Eindrucksurteilen* ist, von Sonderfällen abgesehen, gering. Es wurden knapp mittlere Koeffizienten gefunden. Jedoch ergeben sich fast unter allen Umständen gewisse Übereinstimmungen zwischen verschiedenen Beurteilern, gleichgültig, wie unzureichend die zur Verfügung stehenden Informationen sein mögen.
2. Die *Validität von Eindrucksurteilen* ist verständlicherweise noch geringer, es wurden Koeffizienten zwischen etwa 0,00 und 0.50 gefunden. Auch bei deutlichen Übereinstimmungen zwischen verschiedenen Beurteilern kann die Validität gleich Null sein.
3. Die *Validität von Beurteilungen* steht nur in recht lockerem Zusammenhang mit verschiedenen Persönlichkeitsmerkmalen der Beurteiler. Auch der Zusammenhang zu anderen unabhängigen Variablen ist gering.
4. Der *Inhalt der Beurteilungen* ist enger an andere Bedingungen gebunden als an die Individualität des Beurteilten. Solche Bedingungen sind u.a. die Eigenart des Beurteilers, die Eigenart der sozialen Beziehung zwischen Beurteiler und Beurteiltem und allgemeine Faktoren, wie etwa Stereotype.
5. Das auffälligste *Einzelergebnis* besteht wohl darin, daß die Validität von Beurteilungen weitgehend unabhängig ist von Art und Umfang der Informationen, welche dem Beurteiler über den Beurteilten zur Verfügung stehen."

gründete Diagnostik Entscheidungen und Behandlungszuweisungen ermöglicht, die insgesamt zu nachweislich besseren Ergebnissen führen, als das bei Vernachlässigung dieser Erkenntnisquelle und Anwendung anderer Strategien der Fall wäre.

Aus wissenschaftspragmatischer Sicht ist dabei unerheblich, ob das deterministische Menschenbild, das diesem Ansatz zu Grunde liegt, universell gültig ist oder nicht. Die theoretische Leitvorstellung, daß wenn alle Bedingungen bekannt wären, auch alles vorhergesagt werden könnte, ist, auf menschliches Verhalten angewendet, eine durchaus zweckmäßige Utopie, zumal sie das Auftreten unvorhergesehener Ereignisse keineswegs ausschließt. Ob die perfekte Vorhersage individuellen Verhaltens jemals Realität wird (und ob sie wünschenswert ist), kann vorläufig offen bleiben. Es ist die gemeinsame Aufgabe von Forschung und Praxis zu erkunden, wie weit eine solche heuristische Generaltheorie trägt. Trotz seiner theoretischen und empirischen Unzulänglichkeiten ist dieser Ansatz ein gangbarer und erfolgversprechender Weg, den zu verlassen voreilig wäre, solange es an nachweislich besseren Alternativen fehlt. Für einen solchen Schritt ist die Geschichte der professionellen psychologischen Diagnostik zu kurz.

Diagnostische Feststellungen, z.B. “Thomas ist gar nicht so dumm, wie wir anfangs dachten” oder “Thomas hat einen IQ von 104”, weisen Individuen einer Klasse von Personen zu, die sich in bestimmter Hinsicht untereinander gleichen oder ähneln. In diesem Beispiel wird behauptet, daß Thomas einer Klasse seiner Bevölkerung angehört, die man als durchschnittlich intelligent bezeichnet, unabhängig davon, wie stark sich die Mitglieder dieser Klasse im übrigen, z.B. nach Alter, Geschlecht, sozialer Herkunft oder Lebenswandel, unterscheiden.

Von dem Vorgehen im Alltag hebt sich die professionelle Diagnostik im wesentlichen durch folgende Kriterien ab:

(a) Präzisierung der Begriffe, insbesondere der Merkmale, die erfaßt werden sollen
(b) Präzisierung der Meßoperationen durch
- Standardisierung der Verfahren
- Ökonomisierung der Informationsaufnahme und -Verarbeitung
- Bereitstellung von Vergleichsmaßstäben

(c) Verifizierung der diagnostischen Aussagen und der darauf gestützten Entscheidungen.

Auf diese Kriterien wird in den folgenden Abschnitten näher eingegangen.

1.4.1 Präzisierung der Merkmale

1.4.1.1 Person und Merkmal

In der Persönlichkeitstheorie wird “Person” als eine je einzigartige, unteilbare Ganzheit mit vielschichtigen Bezügen zu sich und ihrer Umwelt verstanden. Psychologisch gesehen, sind Personen hochkomplexe, sich selbst bewußte Systeme mit dem “Ich” als Zentrum des Erlebens und der Verhaltenssteuerung. Aus den bereits dargelegten Gründen könnte der Anspruch, solche individuellen ganzheitlichen Gefüge erschöpfend abzubilden, weder von der professionellen Diagnostik noch von irgendeiner anderen existierenden Diagnostik eingelöst werden. Darauf bezogene Vorbehalte gegenüber der psychologischen Diagnostik sind auch deshalb unerheblich, weil wir uns in der Praxis je nach Fragestellung offenbar ohne zu großen Informationsverlust auf

bestimmte Ausschnitte beschränken können. Unsere Aussagen kennzeichnen nicht die Person schlechthin, sie beziehen sich auf die Ausprägung definierter ***Merkmale,*** die der betreffenden Person zusammen mit einer Vielzahl anderer ***Personen als Merkmalsträger*** zukommen. Dies gilt auch dann, wenn mehrere Einzelmerkmale zu einem Merkmals-Ensemble höherer Ordnung integriert werden (z.B. "Petra ist hochgradig introvertiert").

Mit "Merkmal" ist ein für die Diagnostik zentraler Begriff eingeführt, der uns bereits aus anderen Teilgebieten der Psychologie geläufig ist. Im psychologischen Sprachgebrauch bezieht sich der Begriff "Merkmal" (oder Variable) auf einen ***stabil unterscheidbaren Aspekt mit mindestens zwei Ausprägungsvarianten, anhand derer Objekte gruppiert und Veränderungen an Objekten festgestellt werden können.*** Unter "Objekt" sind hier Personen, Verhaltensprodukte (z.B. schriftliche Klassenarbeiten) und Sachverhalte (z.B. "Hans wächst in einem anregungsarmen häuslichen Milieu auf') zusammengefaßt. Die kaum übersehbare Fülle von Merkmalen ist nach verschiedenen, sich mehrfach überschneidenden Gesichtspunkten zu ordnen. Gerade in der psychologischen Diagnostik müssen wir - aus theoretischen wie aus praktischen Gründen - auf eine solche Ordnung Wert legen, weil diagnostische Aussagen nur dann richtig formuliert und verstanden werden können, wenn Art, Qualität und Funktion der Merkmale, für die Informationen vorliegen, sorgfältig beachtet worden sind. Tabelle 1 (S.24/25) enthält eine schematisierte Übersicht über die wichtigsten Aspekte zur Unterscheidung diagnostisch relevanter Merkmalsklassen.

Als Psychologen sind wir in erster Linie an Merkmalen interessiert, die den einzelnen Personen zukommen, insbesondere natürlich an psychologischen. Doch setzt die operationale Definition mancher psychologischer Merkmale, z.B. des IQ, die Kenntnis nicht-psychologischer Merkmale, vor allem von Zeit- und Altersvariablen, voraus, die dadurch die Funktion von unabhängigen Variablen erhalten. Ebenso kann die Bedeutung individueller diagnostischer Rohwerte durch andere biologische und soziographische Merkmale, wie Geschlecht, Krankheit, besuchter Schultyp, soziale oder ethnische Herkunft, und durch spezifische Umweltvariablen, z.B. Kriminalität des Vaters oder aktuelle Scheidungsauseinandersetzungen der Eltern, relativiert werden.

1.4.1.2 Anlage und Umwelt

Unter ***Umwelt*** verstehen wir die Gesamtheit aller Reize, die während seines Lebens auf ein Individuum wirken, wobei die Reizauswahl und die Reizwirkung durch die vom Individuum ausgehende Aktivität mehr oder weniger mitbestimmt werden. Wir unterscheiden unscharf, aber zweckmäßig, drei Klassen von Umweltvariablen: physikalisch-chemische (z.B. intrauterines Milieu; Klima, Wetter, Luftqualität), materielle (Versorgung mit Bedarfsgütern) und soziokulturelle (personale Beziehungen, mentale Anregungsbedingungen). Ferner unterteilen wir die Umweltvariablen nach der Dauer und Kontinuität ihrer Anwesenheit, wiederum unscharf, in langfristig wirksame und aktuelle (situative). Die Umweltfaktoren (U) sind offensichtlich am Zustandekommen des Verhaltens (V) von Personen beteiligt, doch wissen wir in den meisten Fällen nicht genau, wie das geschieht, welchen Anteil sie daran haben und wie sie mit den genetischen Faktoren (G) zusammenwirken.

Tabelle *1.1:* Vereinfachte Übersicht über diagnostisch bedeutsame Merkmalsklassen.

Klassenbildender Gesichtspunkt	*Merkmalsklassen*	*Beispiele*
Merkmalsträger und Objektbereich	Person-Merkmale - biologische - soziobiographische - psychologische	 Geschlecht, Körpergröße Sozialschichtzugehörigkeit kognitive Entwicklung
	Umwelt-Merkmale - physikalisch-chemische - materielle - soziokulturelle	 Lufttemperatur, Lärmpegel Prokopf-Einkommen der Familie soziales Klima
Datenherkunft	direkt (am Pb) erhobene Merkmale	Schulleistungstestergebnisse, Erfolgs-/Mißerfolgsattribuierung, Verhaltensstile
	indirekt (über den Pb) von Dritten erhobene Merkmale	Auskünfte von Lehrern, Eltern, Behörden (Akten)
Anzahl der unterscheidbaren Ausprägungen	alternativ (qualitativ) dichotomisiert mehrkategorial (qualitativ) kontinuierlich	Geschlecht, rechts/links oberhalb/unterhalb des Medians Berufe. Konfessionszugehörigkeit Alter, Intelligenz, Extraversion
Skalenniveau	nominalskaliert ordinalskaliert	Beruf des Vaters/der Mutter soziale Hierarchien (Hackordnungen), Präferenzen
	intervallskaliert Verhältnisskala	IQ, emotionale Stabilität Alter. Körpergröße
Verteilungsform	normalverteilte Merkmale nicht-notmal-, z.B. asymmetrisch verteilte Merkmale	IQ, Ängstlichkeit, Konzentrationsleistungen Schulnoten in der Grundschule, Fehlerzahlen
Meßgenauigkeit	Merkmale mit hoher innerer Konsistenz	kognitive Leistungsgeschwindigkeit, IQ, Extraversion
	Merkmale mit geringer innerer Konsistenz	Aggressivität im PFT, die meisten Merkmale in Formdeuteverfahren
Dauerhaftigkeit der Merkmalsausprägung	längerfristig stabile Merkmale	Intelligenz Jugendlicher, Introversion/Extraversion
	(periodisch) schwankende Merkmal kurzfristige, aktuelle Merkmale	Verstimmtheit, Wohlbehagen Zustandsangst, Erregtheit, Arger, Müdigkeit, Freude
Komplexitätsgrad	hoch komplexe (mehrdimensionale) Merkmale	Intelligenz, Interessenspektrum, Schulerfolg
	typologisch gebündelte Merkmale	Extraversion/Introversion, Neurotizismus, Maskulinität
	"einfache" (eindimensionale) Merkmale	Sehscharfe, feinmotorische Steuerung, erlebte elterliche Strenge
Generalisierungsbreite	situationsübergreifende Merkmale	emotionale Labilität, Intelligenz, Rigidität
	situationsgebundene Merkmale	spezielle berufliche Fertigkeiten, Phobien, Prüfungsangst, spezielle Einstellungen
Instrumentelle Funktion	abhängige Merkmale (Kriterien) unabhängige Merkmale (Prädiktoren) deskriptive Merkmale	Schulleistung von IQ Schulleistung für Lebenserfolg Schulversagen
	explikativ verwendete Merkmale Moderatorvariablen	soziale Deprivation für Schulversagen Motivation im Zusammenhang von IQ und Schulleistung
Diagnostische Relevanz (relative Validität)	hoch valide Prädiktoren	Intelligenz für Mathematikleistung, Vorkenntnisse
	gering valide Prädiktoren	Selbstbild und Verhaltensstile für Zeugnisnoten

Fortsetzung von Tabelle 1.1:

Klassenbildender Gesichtspunkt	*Merkmalsklassen*	*Beispiele*
Theoretischer Status (diagnostische Aussageebene)	beobachtete Merkmale (Verhalten und Verhaltensprodukte)	Schreibgeschwindigkeit, Lesefehler, Gedächtnisleistungen, Testrohwerte, Aufsätze
	Indikatorvariablen	Zeugnisnoten, IQ
	erschlossene Merkmale (latente Merkmale, Konstrukte, intervenierende Variablen, Dimensionen, Dispositionen)	kognitive Leistungsfähigkeit (g-Faktor), Sprachverständnis, Leistungsmotivation, Eigenschaften wie Zuverlässigkeit, Hilfsbereitschaft

Unsere allgemeine Vorstellung, derzufolge jedes menschliche Verhalten als eine Funktion von Anlage- und Umweltfaktoren zu verstehen ist,

$$V = f(G, U) \qquad [1.1]$$

wird nach dem gegenwärtigen Erkenntnisstand gewöhnlich durch die Modellannahme spezifiziert, daß sich die empirische Varianz (Var) eines psychologischen Merkmals (X) additiv aus genetischen Anteilen (G), Umweltanteilen (U), den Kovarianzen (Cov G,U) zwischen diesen Komponenten sowie den Wechselwirkungs- (G . U) und Meßfehleranteilen (E) zusammensetzt. Unter der Voraussetzung, daß außer dem Meßfehler auch die Wechselwirkungskomponente mit den anderen Anteilen nicht korreliert, ergibt sich folgende Varianzzerlegung:

$$\mathrm{Var}(X) = \mathrm{Var}(G) + \mathrm{Var}(U) + 2\mathrm{Cov}(G,U) + \mathrm{Var}(G.U) + \mathrm{Var}(E). \qquad [1.2]$$

Obwohl diese Gleichung für psychologische Merkmale bis jetzt nicht befriedigend ausgefüllt werden kann, hat sie neben ihrer theoretischen Bedeutung auch eine ***unmittelbare praktische Konsequenz*** für die Diagnostik: Sie verbietet in der großen Mehrzahl der Fälle die "ätiologisch" einseitige Interpretation psychologischer Befunde. Ob, bzw. wieweit z.B. ein festgestellter Rückstand in der Sprachentwicklung bei im übrigen unauffälliger Intelligenz auf das häusliche Milieu des Kindes, auf erworbene organische Mängel oder auf genetische Faktoren zurückgeht, ist den Testwerten nicht zu entnehmen und auch anhand von Anamnesedaten vielfach nicht zu entscheiden.

Doch selbst wenn eine Gleichung vom Typ [1.2] für ein Merkmal in einer Bevölkerung vorläge, wäre zu berücksichtigen, daß die Elemente der Gleichung von Individuum zu Individuum unterschiedliche Werte annehmen können und wir die Zusammensetzung im Einzelfall kennen müßten. Außerdem ist damit noch nicht gesagt, was die aufgeklärten Varianzanteile im Hinblick auf die Möglichkeit bedeuten, das registrierte Verhalten durch gezielte Beeinflussung zu verändern (Merz & Stelzl, 1977).

Selbst wenn wir sicher sein könnten, daß z.B. die kognitive Leistungsfähigkeit, wie sie in Intelligenztests gemessen und im IQ zusammengefaßt wird, im Mittel zu höchstens 20 % durch Umwelteinflüsse determiniert ist, bliebe offen, welche pädagogischen Handlungsspielräume damit eröffnet würden. Die Gleichsetzungen "genetisch bedingt = schwer beeinflußbar" und "umweltbedingt = leicht beeinflußbar" gelten nur eingeschränkt. Organische Mängel, die das Lernen erschweren, können auf genetischen Faktoren beruhen, aber auch die Folge von Umwelteinwirkungen sein. Es kommt hinzu, daß die Wirksamkeit pädagogischer Behandlung vom Lebensalter als

einer wichtigen Moderatorvariablen abhängen kann. Offenbar gibt es auch in der menschlichen Entwicklung so etwas wie "sensible Phasen", in denen stabile Verhaltensmuster leichter als nachher oder vorher erworben werden können (z.B. Urvertrauen, Spracherwerb, Sozialverhalten).

Mit Vorbehalt lassen sich die verschiedenen Verhaltensbereiche allenfalls in eine sehr grobe, komplementäre Rangfolge ihrer Determiniertheit durch Anlage und Reifung bzw. durch Umwelteinflüsse und Lernen bringen. An deren einem Ende mit den höchsten genetischen Anteilen und relativ geringer individueller Variabilität befinden sich die Reflexe und das Instinktverhalten, abnehmend über die Psychomotorik, die Intelligenz und Persönlichkeitsmerkmale i.e.S. bis hin zum anderen Ende mit Einstellungen, Gewohnheiten, Meinungen und z.B. der Konfessionszugehörigkeit, bei denen es unmittelbar einleuchtet, daß hier die Umwelteinflüsse eine deutlich größere Rolle spielen.

1.4.1.3 Kollektiv und Individuum

In diesem Zusammenhang ist an ein weiteres, analoges Problem zu erinnern, das die Diagnostik direkt betrifft. In der Praxis beruhen viele individualdiagnostische Urteile, vor allem Verhaltensvorhersagen, auf ***ideographischen Rückschlüssen*** aus Datenverhältnissen, die theoretisch wie empirisch für Kollektive gelten. Wird z.B. die Empfehlung, ein schulpflichtig gewordenes Kind besser noch nicht einzuschulen, u.a. auf das schwache Abschneiden in einem Schuleingangstest gestützt, ist unbekannt, ob dieses Kind zu denen gehört, für die die Prognose aufgrund der substantiellen Korrelation zwischen Testergebnis und Schulerfolg zutrifft, oder ob es der Gruppe von Kindern angehört, für die sich - aufgrund welcher Randbedingungen auch immer - die Vorhersage nachträglich als falsch erweisen würde. Da wir empirisch nie mit perfekter Abhängigkeit rechnen können, hat dies zur Folge, daß wir grundsätzlich nicht wissen, ob die Vorhersage im Einzelfall zutrifft oder nicht. Die Kennwerte der Kollektive haben lediglich die Funktion von Erwartungswerten für eine Vielzahl von Einzelfallen. Dieses Problem ist bislang nicht befriedigend zu lösen, auch nicht, oder nur sehr bedingt, über individuelle Meßwiederholungen im Sinne der psychometrischen Einzelfalldiagnostik (s. Abschnitt 9.1). Wir müssen uns mit dem Nachweis begnügen, daß wir je nach gegebener Datenlage, insbesondere nach Maßgabe der Korrelation zwischen den Variablen, bei einer größeren Zahl von Urteilen insgesamt weniger Fehler begehen, als dies bei Anwendung anderer verfügbarer Entscheidungs-Strategien der Fall wäre. Dies ist eine der unvermeidlichen Konsequenzen aus der erwähnten Beschränkung unseres Erkenntnisspielraums.

Auch diese allgemeinen Charakteristika psychologischer Merkmalszusammenhänge sind bei der Formulierung diagnostischer Befunde zu beachten. Unabhängig von ihrem Verwertungszweck sind diagnostische Aussagen in der Regel zunächst deskriptive Feststellungen, die sich auf den Ist-Zustand von Merkmalsausprägungen und deren Verknüpfung an Individuen beziehen. Wie präzise unsere Aussagen sein können und welche Schlußfolgerungen sich daraus ziehen lassen, hängt von der Qualität der Daten, d.h. von der Beschaffenheit der benutzten Verfahren und von den Untersuchungsbedingungen ab.

1.4.1.4 Diagnostische Konstrukte

In diesem Zusammenhang wird deutlich, daß psychologische Merkmale nichts unmittelbar Gegebenes sind. Ein Merkmal ist in jedem Fall die ***begriffliche Fassung*** eines Aspekts, in dem sich Individuen voneinander und zu verschiedenen Zeiten unterscheiden können. Das begriffliche Abstraktionsniveau kann dabei zwischen der Wiedergabe eines unmittelbar beobachteten Verhaltens, z.B. sich-Melden im Deutschunterricht am Freitag letzter Woche, und erschlossenen theoretischen Konstrukten, wie "Lerneifer", "Selbstbewußtsein", und "Geltungssucht", variieren. Den unterschiedlichen Aussageebenen entsprechen Unterschiede im theoretischen Status der Aussagen. Je höher das Abstraktionsniveau, desto mehr Implikationen sind darin enthalten.

Konstrukte sind hier als eine Art sprachlicher Kürzel zu verstehen, die inhaltlich definierte Bereiche (grundsätzlich) beobachtbaren Verhaltens von Individuen zusammenfassen. Diagnostische Aussagen auf Konstruktebene stellen individualisierte Dispositionsprädikate dar (Herrmann, 1973). Unter ***Disposition*** wird die Bereitschaft zu bestimmten Handlungen bzw. eine (genügend große) Wahrscheinlichkeit für das Auftreten einer bestimmten Klasse von Verhaltensweisen verstanden. So impliziert z.B. die Feststellung, "Steffi ist hoch leistungsmotiviert", daß Steffi - unabhängig von der Qualität ihrer Leistungen - seit einiger Zeit bemüht ist und voraussichtlich auch weiterhin bemüht sein wird, bei einer Vielzahl verschiedener Leistungsanforderungen jeweils "ihr Bestes zu geben". Solche Aussagen sind nur dann gerechtfertigt, wenn das betreffende Konstrukt, hier die überdauernde Leistungsmotivation, als hinreichend gesichert gelten kann, und wenn sie sich auf Meßoperationen stützen, von denen vorgängig gezeigt worden ist, daß sie ***konstitutiv*** für das Konstrukt sind. Konstrukte gelten in dem Maß als empirisch gesichert, in dem unterscheidbare Ansätze zu ihrer operationalen Realisierung für dieselben Meßwertträger-Kollektive zu konkordanten Ergebnissen führen, im Idealfall, wenn sie sich als "methodeninvariant" erweisen. In der psychologischen Diagnostik ist dies bisher für das Konstrukt "Intelligenz" am vergleichsweise besten gelungen.

Wir unterscheiden demnach die hypothesengeleitete ***induktive Gewinnung*** der Konstrukte in der differentiell-psychologischen und diagnostischen Grundlagenforschung von ihrer ***deduktiven Verwendung*** in der diagnostischen Praxis. Es liegt auf der Hand, daß deren Ergebnisse insgesamt nicht besser ausfallen können, als es dem Konsolidierungs-Status der Konstrukte entspricht. ***Die Sicherung und die sachgerechte Handhabung von Konstrukten ist ein Kernstückprofessioneller Diagnostik.*** Für die Praxis ist es dabei nicht entscheidend, ob ein Konstrukt restlos auf empirisch beobachtetes Verhalten reduziert werden kann. Eine solche Forderung strikt einzulösen, ist offenbar nicht möglich, aber auch nicht nötig. Konstrukte können durchaus einen undeckbaren Bedeutungsüberschuß enthalten, jedenfalls solange und soweit sich mit ihren diagnostischen Ableitungen befriedigende Resultate erzielen lassen, d.h. solange und soweit sich damit Verhaltensvorhersagen nachweislich verbessern lassen.

Es wäre allerdings ein Mißverständnis, wenn man Konstrukte dieser Art als verursachende Instanzen im Sinne latenter Verhaltensdeterminanten interpretieren wollte. Die Aussage "Thomas kann gut denken, weil er überdurchschnittlich intelligent ist", wäre eine bloß schein-kausale Verknüpfung tautologischer Argumente, denn Intelligenz ist durch Denken-Können definiert. Doch ist es vertretbar, Konstrukte in quasi-explikativer Weise zu benutzen. "Da wir bei Thomas (u.a.) einen hohen IQ ermittelt haben, erwarten wir, daß er im Mathematikunterricht gut zurechtkommt". Eine sol-

che Aussage stützt sich auf die Tatsache, daß die Wahrscheinlichkeit des Erfolges im Mathematikunterricht (unter sonst gleichen Bedingungen) deutlich mit dem IQ zunimmt. Auch hierbei greifen wir lediglich auf eine "Wenn-Dann"-Beziehung (⟶) im Sinne eines deskriptiv-empirischen Zusammenhangs zwischen operationalen Repräsentanten der Konstrukte Intelligenz und Schulerfolg zurück; seine Enge bemißt sich an der Höhe eines Korrelationskoeffizienten.

Ungeachtet seiner wissenschaftstheoretischen Belastung stellt das hier skizzierte Verständnis von Konstrukten einen (vorläufig) gangbaren Weg dar, in der Diagnostik Merkmale zu definieren und zu präzisieren, ohne daß dabei das dialektische Verhältnis von theoretischem und operationalem Zugang aus dem Blickfeld gerät. Damit steht uns - unabhängig von ihrer theoretischen und methodischen Ausgestaltung im einzelnen - eine tragfähige allgemeine Basis für eine theoriegeleitete und zugleich handlungsorientierte Diagnostik zur Verfügung.

Ein Beispiel für die Beziehung zwischen Verhaltensmerkmalen und Konstrukten ist in der Abbildung 1 (S. 29) wiedergegeben. Es verdeutlicht die verschiedenen Aussageebenen und den damit zunehmenden Abstraktionsgrad der Merkmale. Zugleich wird die Komplexität von Stufe zu Stufe größer. Dies beruht in diesem Ordnungsansatz auf der empirischen Korrelation zwischen den Merkmalen. Zusammenfassungen dieser Art sind möglich und diagnostisch zweckmäßig, soweit gezeigt werden kann, daß die einbezogenen Merkmale miteinander systematisch höher als mit anderen Merkmalen korrelieren. Das Beispiel orientiert sich an dem deskriptiv-hierarchischen Persönlichkeitsmodell von Eysenck (z.B. 1975) und stützt sich auf Konstruktionsdaten des Persönlichkeitsfragebogens für Kinder (PFK 9-14; Seitz & Rausche, 1976, 1992).

Weitere Möglichkeiten zur Bildung komplexer diagnostischer Klassen bestehen in der Gruppierung von Personen nach der Ähnlichkeit ihrer Merkmalsprofile. Die Gruppen werden dabei so zusammengefaßt, daß die Mitglieder einer Gruppe sich möglichst ähnlich, die Gruppen untereinander möglichst unähnlich sind (Clusteranalyse; s. Abschnitt 4.3.). Dies spielt z.B. bei der Interaktion von Unterrichtsmethode und Schülertyp eine Rolle ("Aptitude-Treatment-Interaction", ATI; vgl. Cronbach & Snow, 1977).

1.4.1.5 Person, Situation und aktuelle Befindlichkeit

Da Gleichungen vom Typ der Formel [1.2] (s.S. 25) vorläufig nicht zu realisieren sind und dies für eine brauchbare Diagnostik auch nicht nötig ist, betrachten wir Verhalten unscharf, aber heuristisch vertretbar, als eine Resultante aus Personmerkmalen (P) und situativen Umweltbedingungen (S):

$$v = f(P, S). \qquad [1.3]$$

Die längerfristigen Umwelteinwirkungen sind hier in den Ist-Zustand der Personmerkmale eingegangen.

Analog zu der in [1.2] angegebenen Varianzzerlegung setzt sich die theoretische Varianz des Verhaltens aus Anteilen zusammen, die auf Unterschiede zwischen den Personen (P) und zwischen den Situationen (S) sowie auf die Kovarianz (Cov P, S) und die Wechselwirkung (**P** · S) dieser Komponenten zurückgehen.

Bei diesem gelegentlich auch "interaktionistisch" genannten Ansatz ist zu beachten, daß die Merkmalsklassen (P) und (S) - abgesehen von ihrer Kovarianz und Wech-

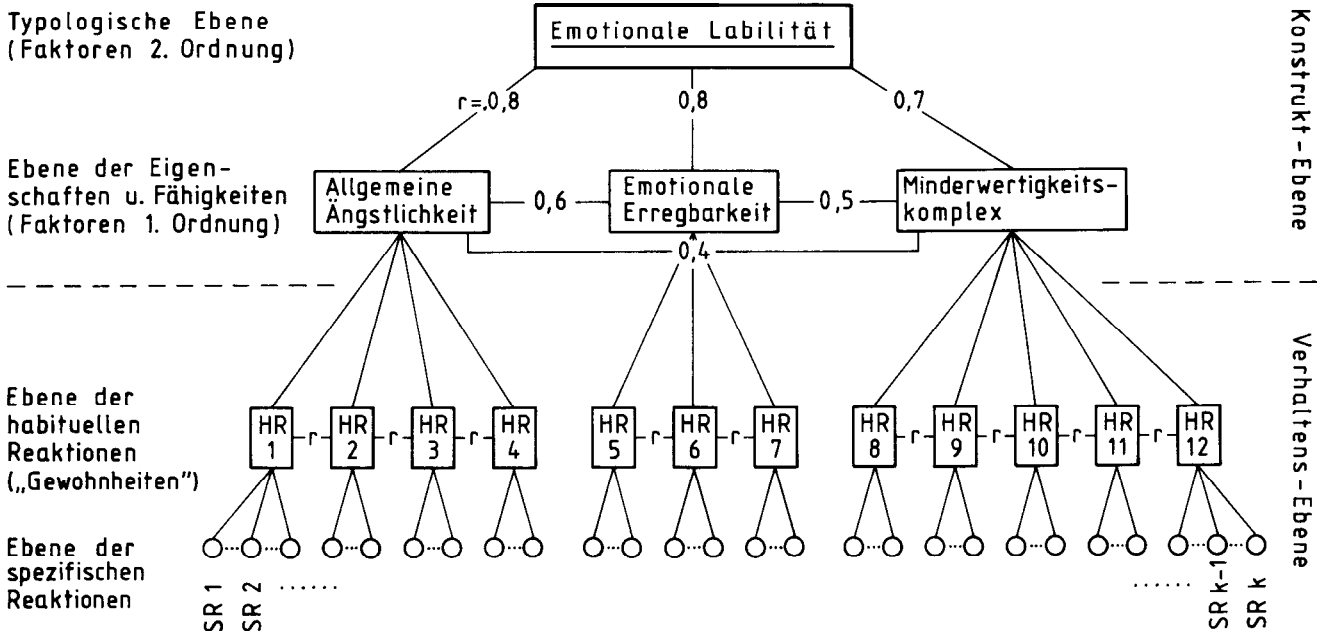


Abbildung 1.1: Konstrukt "Emotionalität" bei Kindern (nach dem deskriptiv-hierarchischen Persönlichkeitsmodell von Eysenck und Konstruktionsdaten des PFK 9-14 von Seitz & Rausche). Bedeutung der habituellen Reaktionen (HR):

HR 1: sich vor körperlicher Verletzung fürchten
HR2: sich vor Dunkelheit und Alleinsein fürchten
HR3: sich vor Bedrohung fürchten
HR4: Angstträume haben
HR5: leicht irritiert und frustriert sein
HR6: unruhig und ungeduldig sein
HR7: nervös sein und zu versagen fürchten
HR 8: andere für hübscher halten
HR 9: andere für fähiger halten
HR 10: sich nach anderen richten
HR 11: sich verspottet fühlen
HR12: darunter leiden, manches nicht zu können

selwirkung - schon von vornherein nicht unabhängig voneinander variieren. Zwar können wir im allgemeinen Ereignisse und Sachverhalte, die außerhalb einer Person vorkommen, eindeutig von Merkmalen unterscheiden, die unmittelbar oder mittelbar einer Person zuzuschreiben sind. Doch enthält "Situation" nach unserer Definition von Umwelt Anteile von Personvarianz. Die Unterscheidung von Person und Umwelt ist theoretisch notwendig; für das praktische Vorgehen bleibt sie fiktiv, weil Person und Umwelt psychologisch nur quasi-unabhängige Merkmalsklassen darstellen.

So wie sich Umwelteinflüsse nach kürzer- oder längerfristigen, situationsspezifischen oder situationsübergreifenden unterscheiden lassen, unterscheiden wir relativ stabile "habituelle" Personenmerkmale (P_{stab}; Eigenschaften, Fähigkeiten, "Persönlichkeitsmerkmale" i.e.S.) von Merkmalen der aktuellen Befindlichkeit (P_{akt}), die die stabileren Merkmale überlagern. Unter der Voraussetzung, daß die Wechselwirkungskomponente mit den übrigen Anteilen nicht korreliert, ist daher die theoretische Varianz des Ausdrucks P in Formel [1.3] definiert als

$$[1.4] \quad Var(P) = Var(P_{stab}) + Var(P_{akt}) + 2Cov(P_{stab}, P_{akt}) + Var(P_{stab}. P_{akt}).$$

Für ein einzelnes Merkmal X gilt, daß es sowohl von konstruktverwandten Merkmalen (im Sinne von Abb. 1; s.S. 29) als auch von konstruktfremden überlagert sein kann, z.B. die kognitive Leistungsfähigkeit von der aktuellen Konfliktbelastung oder vom augenblicklichen Gesundheitszustand. Eine Kovarianz z.B. von Leistungsfähigkeit (stabile Komponente) und Konfliktbelastung (aktuelle Komponente) läge vor, wenn sich die Leistungsschwächeren häufiger in leistungsmindernden Konfliktsituationen befänden als die Leistungsstärkeren. Von Wechselwirkung zwischen Leistungsfähigkeit und Konfliktbelastung wäre zu sprechen, wenn die tatsächlich gezeigte Leistung nur bei den habituell Leistungsschwachen oder deutlich stärker bei ihnen als bei den habituell Leistungsstarken beeinträchtigt wäre.

Damit ist angedeutet, daß wir menschliches Verhalten als ein hochkomplexes Geflecht von Merkmalen zu verstehen haben, dessen systematische Untersuchung Aufgabe der Persönlichkeitsforschung und der Differentiellen Psychologie ist (Amelang & Bartussek, 1990). Weitere Aspekte, durch die sich Merkmale funktional voneinander abheben, sind der Tabelle 1 (s.S. 24/25) zu entnehmen. Darauf wird später noch eingegangen.

1.4.2 Präzisierung der Meßoperationen

1.4.2.1 Standardisierung, Ökonomisierung und Meßgenauigkeit

Die ***grundlegenden diagnostischen Prinzipien*** bestehen darin, durch Standardisierung der Untersuchungsbedingungen ***erstens*** die Situationseinflüsse (S) möglichst konstant zu halten, so daß das registrierte Verhalten theoretisch allein als Funktion der Personmerkmale (P) verstanden werden kann. Für S = konstant gilt:

$$[1.5] \quad v = f(P).$$

Zweitens wird durch eine repräsentative Auswahl der Aufgaben, die den Pbn gestellt werden, der Reaktionsbereich inhaltlich so spezifiert, daß sich die registrierte

Verhaltensstichprobe einem bestimmten Merkmal(skomplex) X_j zuordnen läßt. Für S = konstant und inhaltlich spezifiziert gilt:

[1.6] $$V = f(X_j),$$

wobei (X_j) Element der Menge (P) aller Merkmale ist, die wir an Personen unterscheiden können.

Das erste Standardisierungsprinzip entspricht der Forderung nach ***Objektivität*** des Vorgehens. Die Verfahren sollen vor allem durch eindeutige, verbindliche Vorschriften ("Instruktionen") für die Durchführung und die Auswertung sowie für die Interpretation der Ergebnisse gewährleisten, daß der diagnostische Befund so ***wenig wie möglich*** von den ***äußeren*** Umständen abhängt, unter denen die Untersuchung stattgefunden hat. Deren Anteil an der empirischen Varianz der Meßwerte soll gegen null gehen. Zu den äußeren Umständen zählen auch Untersucher und Auswerter (vgl. Abschnitt 2.2).

Auf dem zweiten Standardisierungsprinzip beruht die ***Validität*** (diagnostische Gültigkeit) der Methoden. Sie liegt empirisch zunächst in dem Maße vor, wie es bei der Konstruktion eines Verfahrens gelingt, den theoretisch bestimmten Merkmalsbereich, auf den es sich richten soll, tatsächlich abzudecken. Dies gilt für Klassen- oder Prüfungsarbeiten grundsätzlich ebenso wie für formelle Testverfahren. Unsere Erhebungsinstrumente sind jeweils ***operationale Definitionen*** der Merkmale, die uns diagnostisch interessieren. So kann z.B. die Intelligenz nach dieser oder jener ***Intelligenztheorie*** als der Grad der kognitiven Leistungsfähigkeit eines Individuums definiert werden; ***diagnostisch ist sie durch das definiert, was der Test mißt*** (Boring, 1923, zit. nach Conrad, 1983, S. 107; vgl. Abschnitte 1.4.3 und 2.2). Mit der Standardisierung dienen die beiden genannten Prinzipien zugleich der ***Ökonomisierung;*** d.h. in der professionellen Diagnostik werden die Daten nicht gelegentlich oder zufällig, sondern unter möglichst ***einheitlichen Bedingungen*** möglichst ***systematisch*** und ***treffsicher*** erhoben. Wie beim Experiment geht es um künstlich herbeigeführte, kontrollierte und grundsätzlich wiederholbare Verhaltensbeobachtung; anders als beim Experiment mit seinem Prinzip der Bedingungsvariation gilt hier der Grundsatz der ***Bedingungskonstanz.***

Je ähnlicher ("homogener") die Anforderungen der Aufgaben oder Fragen und je geringer der Einfluß der äußeren Umstände, desto genauer ("reliabler") wird das betreffende Personmerkmal gemessen. Die Varianz der Meßergebnisse soll möglichst vollständig auf systematische Unterschiede in der Merkmalsausprägung bei den untersuchten Personen ("Meßwertträgem") zurückgehen.

In dem Maße, in dem die so entstandene Reihung der Personen bei einer späteren Meßwiederholung erhalten bleibt, kann man von einem "stabilen" Merkmal sprechen. Ein Merkmal ist theoretisch umso stabiler, je näher die Korrelation zwischen erster und zweiter Messung an die Meßgenauigkeit des Verfahrens herankommt und je länger die Messungen auseinanderliegen. Empirisch bleibt die feststellbare Stabilität eines Merkmals auf die Meßgenauigkeit des diagnostischen Verfahrens beschränkt. "Stabil" bezieht sich hier auf die psychologische oder pädagogische Bedeutung des Merkmals, unabhängig davon, wie stark sich inzwischen die Meßwerte ***absolut*** verändert haben. Dies spielt z.B. eine Rolle, wenn es um die Messung von Lernfortschritten geht und daraus auf unterschiedliche "Fähigkeiten" der Schüler geschlossen werden soll. Schlüsse dieser Art werden in der Regel gezogen, wenn ***längerfristige*** Vorhersagen ("Prognosen") gefordert sind, z.B. bei der Einschulungs- oder Eignungs-

diagnostik. Dabei geht man auf die Konstruktebene über, weil Vorherzusagendes Merkmal ("Kriterium") und Ausgangsmerkmal ("Prädiktor") phänotypisch verschieden sein können, wie das z.B. bei der Verwendung von Intelligenztests für die Diagnose der "Sonderschulbedürftigkeit" oder die Prognose des Berufserfolgs der Fall ist. Die Stabilität der Merkmale und der Randbedingungen ist Voraussetzung für Vorhersagen. Die erreichbare Güte einer Vorhersage wird begrenzt durch das empirisch ermittelte Maß an Stabilität der beteiligten Merkmale.

1.4.2.2 Vergleichsmaßstäbe

Diagnostische Aussagen kennzeichnen die individuelle Ausprägung eines Merkmals an einer Person. Angenommen, die Schülerin Gabi habe bei einer Reihe von Rechtschreibaufgaben 16 Punkte erhalten. Wie ist diese Leistung zu beurteilen? Die bloße Mitteilung eines solchen Wertes läßt nicht erkennen, was er bedeutet. Dazu sind offenbar weitere Informationen nötig, z.B. welches ist die höchste erreichbare oder die höchste erreichte Punktzahl? Wie ist die Skala definiert? Darf man annehmen, daß jemand mit 8 Punkten nur "halb so gut" in Rechtschreiben ist? (Oder "doppelt so gut", falls nämlich Fehler gezählt wurden?) Von der Definition und der Qualität der Skala abgesehen (s. Tabelle l), brauchen wir Bezugsgrößen, die es gestatten, den individuellen Wert auf dem angenommenen Merkmalskontinuum zu lokalisieren. Dies kann auf verschiedene Weise geschehen. Wir können den Wert z.B. auf die Verteilung aller Werte beziehen, die von Schülern desselben Alters oder derselben Schulstufe erreicht werden, und angeben, ob er über oder unter einem ausgezeichneten Kennwert, etwa dem Median oder einer anderen Marke, liegt, und wie weit er davon entfernt ist. Unabhängig davon können wir uns u.U. mit der Feststellung begnügen, ob z.B. vor Beginn einer neuen Unterrichtseinheit die nötigen Mindestanforderungen bei den Schülern erfüllt sind, oder ob ein bestimmter Sollwert erreicht ist, der uns erlaubt anzunehmen, daß die Schüler die betreffende Fertigkeit inzwischen hinreichend sicher beherrschen.

Ganz gleich, ob ein erhobener Istwert für pädagogisch befriedigend gehalten oder als veränderungsbedürftig betrachtet wird, in jedem Fall bedarf es dazu verläßlicher Orientierungsgrößen ("Normen"). Anders wäre die Bedeutung einer diagnostischen Information nicht einzuschätzen; sie bliebe wertlos. Dies gilt erst recht, wenn z.B. ermittelt werden soll, wie "beträchtlich" die Ausfälle im Leistungsspektrum eines Schülers sind, und zu klären ist, ob sie auf mangelhafte "Beschulung", auf starke psychische Belastung (z.B. durch Ängste) oder auf Motivationsstörungen zurückgehen, bzw. ob der Schüler als "lernbehindert" im Sinne von "sonderschulbedürftig" gelten muß. Ebenso benötigen wir Normen, wenn es um das Erkennen besonderer "Begabungen" oder um die Feststellung geht, die Leistungen eines Schülers seien "durchschnittlich" und sein Verhalten "unauffällig": ganz allgemein, wenn Leistungen und Verhalten intra- oder interindividuell verglichen werden sollen, sei es zu einem bestimmten Zeitpunkt, sei es, daß uns Veränderungen von einem zum anderen Zeitpunkt interessieren. Als Orientierungsgrößen können gesetzte ***Sollvorgaben*** ("Gabi hat in Mathematik das Klassenziel nicht erreicht"), individuelle ***Bezugsnormen*** ("Gabi hat in Deutsch-Schriftlich erhebliche Fortschritte gemacht") oder ***Gruppennormen*** dienen ("Im Englischen gehört Gabi zum besten Viertel ihres Jahrgangs"). Häufig ist es zweckmäßig, für Vergleiche zwischen Individuen und Merkmalen einheitliche Skalen zu benutzen.

Normen und die Handhabung von Normen, zumal in der Pädagogisch-psychologischen Diagnostik, sind nichts von Natur aus Gegebenes. Sie hängen von kulturellen und gesellschaftlichen Bedingungen ab und beruhen größtenteils auf Vereinbarungen, z.B. darüber, was, wann und wie in Schulen unterrichtet werden soll, oder welche Zulassungsbedingungen für den Besuch von Sonderschulen oder Universitäten gelten. Dementsprechend unterliegen sie dem Wandel, und sie sind grundsätzlich revidierbar. Dies trifft selbst für das zu, was man in einer Gesellschaft-ungeachtet interkultureller Gemeinsamkeiten - unter Intelligenz versteht. Normen können mehr oder weniger engmaschig sein. Die bei uns übliche Skala für Schulnoten läßt fünf oder sechs, mit den manchmal vergegebenen Zwischennoten zehn bis zwölf Abstufungen zu. Intelligenzquotienten (IQ) sind zwei- bzw. dreistellig definiert und erwecken den Eindruck, man könne in der Gesamtbevölkerung mindestens 90 Ausprägungsgrade der kognitiven Leistungsfähigkeit unterscheiden (IQ zwischen 55 und 145). Wie eng das Raster von Normen sein darf, hängt in erster Linie von der Genauigkeit ab, mit der das abgebildete Merkmal gemessen wird, also von der Reliabilität des diagnostischen Verfahrens, bzw. der Stabilität des Merkmals. Sind diese gering, kann selbst eine einstellige Normenskala eine Differenzierung vortäuschen, die wegen der mangelnden Qualität der Meßoperation nicht gerechtfertigt ist.

In der Praxis kommt es vielfach nicht auf maximale Differenzierung, bzw. die maximal mögliche Meßgenauigkeit an. Häufig genügen Unterscheidungen wie “versetzt”/“nicht versetzt”, “durchschnittlich”, bzw. “unter-” oder “überdurchschnittlich” oder die Feststellung, daß die große Mehrheit der Schüler das Unterrichtsziel erreicht hat, ohne daß für jeden einzelnen nachgeprüft werden müßte, in welchem Maße sich seine Leistungen z.B. von denen der Mitschüler unterscheiden. Andrerseits wird - im Zusammenhang mit der Handhabung von Grundgesetzartikeln, die die Freiheit der Berufswahl garantieren-für die Abschlußzeugnisse von Gymnasien eine ausgefeilte Arithmetik vorgeschrieben. Sie soll die schulische Gesamtleistung jedes Schülers auf einem Raster von 31 zulässigen Skalenwerten lokalisieren und damit eine feine Abstufung der kritischen Mindestwerte für die Zulassung zu bestimmten Studiengängen ermöglichen. Auch hier ist u.a. zu fragen, ob die Meßgenauigkeit der Diagnostik ausreicht, um die “Befunde” so stark zu differenzieren. Ist die Leistungsfähigkeit von Bewerbern, die z.B. mit der Note 1,9 den kritischen Wert nicht erreichen, tatsächlich geringer als die der anderen, die mit 1,8 zugelassen werden?

Allgemein gilt jedoch der Grundsatz, daß Normen pädagogisch umso ergiebiger genutzt werden können, je stärker sie zwischen den Ausprägungsgraden eines Merkmals zu differenzieren gestatten, vorausgesetzt, die diagnostischen Verfahren, auf denen sie beruhen, sind entsprechend meßgenau. Dies interessiert uns natürlich nicht bei beliebigen Merkmalen sondern nur bei solchen, von denen gezeigt werden kann, daß sie für Erziehung und Unterricht bedeutungsvoll sind, und worin diese Bedeutung besteht.

1.4.3 Verifizierung diagnostischer Aussagen

Diagnostische Aussagen beschreiben die individuelle Ausprägung von Merkmalen, auf denen sich Personen unterscheiden können. Wie alle wissenschaftlichen Aussagen über empirische Sachverhalte müssen die Aussagen der professionellen Diagnostik überprüfbar sein. Sie sollen nicht nur objektiv und hinreichend präzise sein, sie

müssen sich auch bewähren, d.h. sie müssen nachweislich und möglichst vollständig zutreffen. Erst damit wird die Diagnostik ihrer Funktion gerecht, zur Optimierung pädagogischer Entscheidungen beizutragen. Die diagnostischen Verfahren sind also darauf zu untersuchen, wieweit sie diesem Anspruch genügen. Auf die Verfahren bezogen, sprechen wir - wie bereits erwähnt - von deren ***Validität*** oder ***Gültigkeit.*** Damit ist das Ausmaß gemeint, in dem etwa ein Test für den Zweck, zu dem er verwendet werden soll, tatsächlich brauchbar ist; z.B. wie gut ein Schuleingangstest als Prädiktor das Kriterium Schulerfolg vorherzusagen gestattet, wenn die Kinder ihren Lernvoraussetzungen entsprechend gefördert werden.

An die Validität der Verfahren sind umso höhere Ansprüche zu stellen, je gewichtiger die zu treffende Entscheidung ist. Wo es entsprechend gute Verfahren (noch) nicht gibt, muß die daraus resultierende Unsicherheit berücksichtigt werden; d.h. die Randbedingungen und die wahrscheinlichen Konsequenzen alternativer Entscheidungen sind so sorgfältig wie möglich abzuwägen. Bleibt die empirische Fehlerquote auch bei Nutzung aller verfügbarer Prädiktoren hoch, sind u.U. die systembedingten Entscheidungszwänge zu revidieren. Dies betrifft z.B. die vom traditionellen westdeutschen Schulsystem geforderten Übergangsentscheidungen nach dem vierten Grundschuljahr (vgl. Tent, 1969). Der im vorigen Abschnitt angeführte Zugang zum Studium von Numerus-Clausus-Fächern ist ein anderes Beispiel für die Frage nach der empirischen Legitimation staatlicher Regelungsbefugnisse. Hierbei geht es hauptsächlich um einen Aspekt der Validität von Lehrerurteilen. Den Inhabern von Reifezeugnissen wird die unbefristete Eignung und Berechtigung bescheinigt, beliebige Fächer an wissenschaftlichen Hochschulen studieren zu können. Wieweit ist die scheinbar plausible Annahme gerechtfertigt, daß sich Abiturienten für bestimmte, zulassungsbeschränkte Studiengänge umso eher eignen, ***je*** besser die ***Durchschnittsnote*** ihres Zeugnisses ist? Und haben Absolventen mit besseren Prüfungsergebnissen auch mehr Erfolg im Beruf?

In diesen Beispielen ergibt sich die Validität der diagnostischen Verfahren aus dem Verwertungszusammenhang. Wir sprechen dann von ***Kriteriumsvalidität*** und von ***prognostischer Validität.*** Unter den Rahmenbedingungen unseres Schulsystems spielt dieser Validitätsaspekt eine unverhältnismäßig große, wenn auch inzwischen abnehmende Rolle.

Die Verfahren können aber unabhängig von ihrer aktuellen Verwertung auf ihre pädagogische oder psychologische Bedeutung überprüft werden. Welches Merkmal, oder welche Merkmalskombination, wird erfaßt? Geben z.B. die Deutschnoten tatsächlich nur die Leistung der Schüler im Deutschunterricht wieder oder gehen vielleicht das "Betragen" oder die Sympathie/Antipathie auf seiten des Lehrers mit ein? Welche Komponenten der kognitiven Leistungsfähigkeit sind in einem Intelligenztest berücksichtigt, und wie groß sind eventuell die Anteile von Motivation und Konzentration? Wieweit beeinflußt die Neigung, sozial erwünscht zu reagieren, die Ergebnisse eines Angstinventars oder Persönlichkeitsfragebogens? Dabei interessiert in erster Linie, wieweit die empirischen Daten mit theoretisch vorgegebenen Merkmalskonzepten (wie "Intelligenz" oder "Labilität") in Einklang stehen. Man spricht in diesen Fällen (unscharf) von ***Konstruktvalidität.*** Was ein konstruktvalider Test mißt, kann je nach diagnostischer Fragestellung und Vergleichsgröße (Kriterium) verschieden belangvoll sein. Ein Intelligenztest, der z.B. hoch mit der objektiven Schulleistung in Mathematik korreliert, kann u.U. Zeugnisnoten nur mäßig genau und die Ergebnisse mündlicher Prüfungen noch weniger genau vorhersagen. Obwohl alle drei

Variablen mit Intelligenz zu tun haben, ist das Ausmaß unterschiedlich; dieses hängt u.a. von der instrumentellen Qualität des Kriteriums ab. Es ist also nicht sinnvoll, von ***der*** Validität eines Verfahrens zu sprechen; vielmehr gibt es je nach Verwendungszweck eine Mehrzahl unterscheidungsbedürftiger ***Validitätsaspekte.***

In der Pädagogisch-psychologischen Diagnostik spielt darüber hinaus die ***Lehrplangültigkeit*** oder ***curriculare Validität*** eine besondere Rolle. Hier geht es um den (meist über Expertenurteile erbrachten) Nachweis, daß die Aufgaben in Schulleistungstests für die Lehrplananforderungen eines zeitlichen Ausschnitts aus einem Unterrichtsfach repräsentativ sind. So muß z.B. ein curricular valider Rechentest für das vierte Schuljahr genau die Typen von Aufgaben enthalten, die vom Lehrplan für den Mathematikunterricht auf dieser Schulstufe vorgesehen sind, also schriftliches Multiplizieren und Dividieren, Kopfrechnen und Textaufgaben ("Rechnerisches Denken" in Sachzusammenhängen). Aber auch jede Klassenarbeit muß selbstverständlich, meist für einen kleineren Ausschnitt, "lehrplangültig" sein. Zu beachten ist, daß die Leistungen der Schüler immer auch von der Güte des erteilten Unterrichts mitbestimmt werden. Die ***Rückmeldungsfunktion*** solcher diagnostischer Erhebungen gilt gleichermaßen der Schule wie den Schülern.

Ganz gleich wie die Validität eines diagnostischen Verfahrens bestimmt wird, es geht jeweils um die Aufklärung der Varianz uns interessierender Personmerkmale, und zwar unabhängig davon, ob wir es mit relativ stabilen oder weniger stabilen Merkmalen zu tun haben. Die Validität ist das wichtigste Gütekriterium aller Diagnostik. Unsere Aussagen sollen so valide sein wie möglich, d.h. die Unterschiede, die wir feststellen, sollen so genau wie möglich zutreffen. Doch können auch weniger valide Verfahren nützlich sein. Ihre Anwendung ist gerechtfertigt, solange keine nachweislich besseren zur Verfügung stehen. Wie schon angeführt, müssen wir stets bedenken, ***wie valide*** ein Verfahren für den Zweck ist, zu dem wir es benutzen.

1.5 Zusammenfassung und Definition von Diagnostik

Verhalten, Leistungen, Eigenschaften und Fähigkeiten von Personen zu beurteilen, ist uns aus dem alltäglichen zwischenmenschlichen Umgang von früh an vertraut. Unsere Urteile zielen darauf ab, den anderen möglichst gut zu verstehen und einzuschätzen, was wir von ihm zu erwarten haben. Dies ist Teil unserer Bemühungen, mit Hilfe bestandsfester Erkenntnisse die Lebenswelt, in der wir agieren, überschaubar zu machen und künftige Ereignisse weniger ungewiß erscheinen zu lassen. Aufgabe der Pädagogisch-psychologischen Diagnostik ist es, dies für den Lebensbereich zu leisten, den wir Erziehung nennen. Erziehen heißt, Merkmale von Personen über mentale Beeinflussung möglichst dauerhaft zu verändern. Differenzierung und Individualisierung sind anerkannte Grundsätze pädagogischen Vorgehens.

Die professionelle Diagnostik dient der Verwirklichung dieser Grundsätze. Sie folgt damit dem allgemeinen Optimierungsgebot, das auch für pädagogisches Handeln gilt. Dabei knüpft sie an die Alltagsdiagnostik an. Ihr wissenschaftliches Fundament erhält sie durch die Klärung ihrer persönlichkeits- und meßtheoretischen Annahmen, durch die Präzisierung der Merkmale, auf die sie sich richtet, durch die genaue Analyse der Randbedingungen, unter denen sie abläuft, durch die Standardisierung und Präzisierung der Meßoperationen (der diagnostischen Erhebungsmethoden),

durch die Bereitstellung von Maßstäben zur Beurteilung der individuellen Meßergebnisse sowie durch die empirische Verifizierung ihrer diagnostischen Aussagen.

Als ***Praxis*** ist Diagnostik in der Regel "problemlösendes Handeln" im Sinne der Anwendung einer grundsätzlich "nutzenmaximierenden Technologie auf wissenschaftlicher Grundlage" (Wottawa & Hossiep, 1987). Im Umfeld der Erziehung geht es primär um einen jeweils pädagogisch definierten Nutzen, worin dieser auch immer bestehen mag. Je nach Fragestellung sind demzufolge Methoden zu verwenden, die den Ansprüchen einer wissenschaftlich fundierten Diagnostik genügen und deren Güte der Tragweite der pädagogischen Schlußfolgerungen entspricht, die man darauf stützen will.

Damit ist der Sache nach deutlich, was unter Pädagogisch-psychologischer Diagnostik zu verstehen ist, wozu sie dient, wie sie vorgeht, und was wir von ihr erwarten. In der Literatur finden sich zahlreiche, unterschiedlich genaue und umfassende Begriffsbestimmungen (z.B. Klauer, 1978; Michel & Conrad, 1982; Ingenkamp, 1985; Jäger & Petermann, 1992). Es erscheint uns zweckmäßig, Diagnostik, in Anlehnung an Tent & Waldow (1984, S. 5) zusammenfassend wie folgt zu definieren:

Definition für Diagnostik

"Diagnostik ist ein theoretisch begründetes System von Regeln und Methoden zur Gewinnung und Analyse von Kennwerten für inter- und intraindividuelle Merkmalsunterschiede an Personen."

Dazu gehören

(a) die Formulierung diagnostischer Probleme und Fragestellungen
(b) die Erhebung diagnostischer Daten und deren Integration ***zu Diagnosen*** sowie
(c) die damit verknüpften Folgeerwartungen ***(Prognosen)*** im Hinblick auf verfügbare oder wünschbare Behandlungsalternativen.

Bei den Erhebungsmethoden unterscheidet man die informelle, instrumentell meist schwächere Urteilsbildung durch Experten (z.B. Lehrer, Psychologen, Ärzte) aufgrund Verhaltensbeobachtung, Leistungseinschätzung und Gesprächsführung von der formalisierten Urteilsbildung mit Hilfe standardisierter Untersuchungsverfahren (Inventarien und ***Tests).***

Mit dieser Definition sind die in Praxis und Forschung möglichen Fälle der Anwendung diagnostischer Prozeduren und der Verwertung diagnostischer Informationen erschöpfend abgedeckt. Wie alle empirisch-psychologischen Untersuchungen werden diagnostische Erhebungen stets an Individuen vorgenommen; diagnostische Aussagen beziehen sich daher primär auf Einzelpersonen, denen damit bestimmte Attribute zugeschrieben werden. Aus den individuellen Ergebnissen lassen sich je nach der Skalenqualität Kennwerte für Gruppen errechnen, so daß man - vor allem zu Forschungszwecken - z.B. Schulklassen, Schultypen und Schulstufen, Schülerkohorten oder Statusgruppen hinsichtlich bestimmter Merkmale insgesamt kennzeichnen und miteinander vergleichen kann.

Von ***Diagnose*** sprechen wir in diesem Zusammenhang, wenn Personen anhand relevanter und valider Einzelinformationen innerhalb eines pädagogisch bedeutsamen Klassifikationssystems einer bestimmten Klasse von Merkmalsträgem zugeordnet werden. So kann z.B. die Kombination des Rohwerts auf einem kognitiven Leistungstest mit dem Lebensalter bei einem jüngeren Kind bedeuten, daß es "überdurchschnittlich", derselbe Rohwert bei einem älteren, daß es "unterdurchschnittlich intelligent" ist. Ähnlich fassen wir verschiedene Informationen über die Sinnestüchtigkeit, die Schulleistung, die Intelligenz und die Vorgeschichte eines Schülers zu Diagnosen wie "lese-rechtschreibschwach" oder "lernbehindert" zusammen.

Unter ***Prognose*** versteht man die Erwartung ("Vorhersage") künftigen Verhaltens oder künftiger Leistungen aufgrund diagnostischer Erkenntnisse. Bezieht sich die Erwartung auf dasselbe Merkmal wie das zuvor diagnostisch erfaßte, ist die Treffsicherheit der Vorhersage eine Funktion der Stabilität des Merkmals; bezieht sich die Erwartung auf ein anderes Merkmal ("Kriterium"), hängt die Treffsicherheit neben der Stabilität des Prädiktors und des Kriteriums von der Enge des empirischen Zusammenhangs zwischen beiden ab. Im pädagogischen Alltag spielen Erwartungen dieser Art eine große Rolle; formalisierte Vorhersagen werden allerdings nur selten genutzt.

Als ***diagnostischen Test*** bezeichnen wir jedes systematisch konstruierte, routinemäßig anwendbare, standardisierte und normierte ***Verfahren zur Erhebung individueller Reaktionsstichproben,*** sofern dessen Meßgüte bekannt ist und für den Verwendungszweck ausreicht.- Diese strenge Bestimmung soll im Sinne unserer Diagnostik-Definition die Unterscheidung "weicher" von methodisch anspruchsvollen Verfahren gewährleisten und dem Verschleiß des Test-Begriffs entgegenwirken. Auch wenn sie methodisch hohen Ansprüchen genügen, sind diagnostische Ergebnisse stets deskriptive Aussagen über Ist-Zustände. Für sich genommen, besagen sie in der Regel noch nichts über die zugrundeliegenden "Ursachen". Dazu bedarf es zusätzlicher Analysen. Ebenso wenig ist diagnostischen Aussagen zu entnehmen, weshalb und wie der festgestellte Ist-Zustand geändert werden soll. Für unseren Anwendungsbereich wird dies von den pädagogischen Zielvorgaben und den Möglichkeiten zu ihrer Realisierung bestimmt ("Primat der Didaktik", Tent & Waldow, 1984; Schlee, 1985). Da Erziehung Veränderungen an Personmerkmalen bewirken soll, kommt der ***Veränderungsmessung*** ("Verlaufsdiagnostik") in der Pädagogisch-psychologischen wie in der klinischen Diagnostik eine besondere Bedeutung zu (s. Abschnitt 9).

Grundlegende Literatur:

Erziehungswissenschaftliche Grundlagen:

Brezinka, W. (1978). *Metatheorie der Erziehung* (4. Aufl.). München: Reinhardt.

Sauer, K. (1981). *Einführung in die Theorie der Schule.* Darmstadt: Wiss. Buchgesellschaft.

Wilhelm, Th. (1977). *Pädagogik der Gegenwart (5.* Aufl.). Stuttgart: Kröner.

Zur Differentiellen Psychologie und Persönlichkeitsforschung:

Amelang, M. & Bartussek, D. (1990). *Differentielle Psychologie und Persönlichkeitsforschung* (3. Aufl.). Stuttgart: Kohlhammer.

Herrmann, Th. (1991). *Lehrbuch der empirischen Persönlichkeitsforschung* (6. Aufl.). Göttingen: Hogrefe.
Hofstätter, PR. (1977). *Persönlichkeitsforschung* (2. Aufl.). Stuttgart: Kröner.
Mogel, H. (1990). *Umwelt und Persönlichkeit. Bausteine einer psychologischen Umwelttheorie.* Göttingen: Hogrefe.

Zur Psychologischen und Pädagogischen Diagnostik:

Ingenkamp, K. (1985). *Lehrbuch der Pädagogischen Diagnostik* (Studienausgabe 1988). Weinheim: Beltz.
Jäger, R.S. & Petermann, F. (Hrsg.) (1992). *Psychologische Diagnostik. Ein Lehrbuch* (2., veränd. Aufl.). Weinheim: Psychologie Verlags Union.
Klauer, K.J. (Hrsg.) (1978). *Handbuch der Pädagogischen Diagnostik,* Band 1. Düsseldorf: Schwann.
Kleber, E.W. (1992). *Diagnostik in pädagogischen Handlungsfeldern.* Weinheim: Juventa.
Süllwold, F. (1983). Pädagogische Diagnostik. In K.J. Groffmann & L. Michel (Hrsg.), *Intelligenz- und Leistungsdiagnostik (S.* 307-386). Göttingen: Hogrefe.
Wottawa, H. & Hossiep, R. (1987). *Grundlagen psychologischer Diagnostik.* Göttingen: Hogrefe.

Weiterführende Literatur zur Pädagogisch-psychologischen Diagnostik:

Ingenkamp, K. (1990). *Pädagogische Diagnostik in Deutschland 1885-1932.* Weinheim: Deutscher Studien Verlag.
Laux, H. (1990). *Pädagogische Diagnostik im Nationalsozialismus 1933-1945.* Weinheim: Deutscher Studien Verlag.

2. Grundzüge der klassischen Testtheorie

1. Von welchen Definitionen und Annahmen geht die klassische Testtheorie aus?
2. Welche Kriterien stellt die klassische Testtheorie zur Verfügung, um die Qualität eines Tests zu beurteilen?
3. Warum hängen die Gütekriterien nicht nur vom Test, sondern auch von der Personenstichprobe ab, an der sie erhoben wurden? Weshalb wird häufig eine Normalverteilung der Testwerte angestrebt?
4. Was sind Testnormen und wozu werden sie verwendet?

Vorstrukturierende Lesehilfe

Die klassische Testtheorie, ihre Grundbegriffe und ihre Gütekriterien für psychologische Tests gehören zum selbstverständlichen Methodenrepertoire psychologischer Diagnostik. Sie hat mit ihren Forderungen nach Objektivität, Reliabilität und Validität die Testentwicklung nachhaltig beeinflußt, und es ist heute kaum mehr vorstellbar, einen Test zu publizieren, ohne zu diesen grundlegenden Gütekriterien Angaben zu machen. Im folgenden sollen die Grundgedanken der klassischen Testtheorie kurz zusammengefaßt werden. Diese Zusammenfassung kann kein Ersatz für eine systematische Einführung sein, wie sie in klassischen Lehrbüchern, z.B. Lord & Novick (1968) oder Fischer (1974) gegeben wird. Auf Formeln wird hier weitgehend, auf Ableitungen ganz verzichtet.

Im folgenden wird zunächst das Konzept des wahren Werts und des Meßfehlers im Sinn der klassischen Testtheorie eingeführt (2.1). Darauf aufbauend werden die Gütekriterien Reliabilität, Validität und Objektivität begrifflich erläutert (2.2). Es wird darauf hingewiesen, daß die für die Gütekriterien errechneten Kennwerte nur in Hinblick auf die Personenpopulation, an der sie bestimmt wurden, zu interpretieren sind (2.3) und die Rolle der Normalverteilung diskutiert (2.4). Der letzte Abschnitt (2.5) behandelt die Bedeutung der Testnormen als Interpretationshilfe, vor allem in der individuell beratenden Diagnostik. Auf eine Gesamtzusammenfassung von Kapitel 2 wird verzichtet, da der Text selbst nicht mehr als eine knapp gehaltene Zusammenfassung einiger Grundbegriffe enthält.

2.1 Grundbegriffe der Klassischen Testtheorie: Beobachteter Wert, wahrer Wert, Meßfehler

Wenn von psychologischer "Testtheorie" die Rede ist, so legt das zunächst die Vermutung nahe, es handle sich um eine Theorie, die nur auf psychologische und päda-

gogische Tests anzuwenden sei. Das trifft jedoch nicht zu. In der klassischen Testtheorie geht es um die allgemeine Frage, wie Gütemaßstäbe für psychologische und pädagogische Messungen zu definieren sind und wie diese Gütekriterien praktisch zu bestimmen sind. Fragen dieser Art sind z.B.: “Wie genau ist die Messung?” “Wie stark wird sie durch zufällige Fehler beeinflußt?” “Wird dasjenige Merkmal gemessen, das gemessen werden soll, oder wird die Messung stark von anderen Merkmalen mitbeeinflußt?” - Solche Fragen stellen sich bei jeder Messung (mittels Tests, Lehrereinschätzungen, Noten, Selbst- und Fremdbeurteilungen usw.), so daß die klassische Testtheorie als ein allgemeiner begrifflicher Rahmen anzusehen ist, der es ermöglicht, die Qualität von Messungen zu beurteilen und die Auswirkungen von Meßfehlern abzuschätzen. Wenn im folgenden auch meistens von “Tests” die Rede sein wird, so sind Begriffe und Aussagen leicht auf andere Arten von Messungen zu übertragen.

In der klassischen Testtheorie geht man davon aus, daß die Ergebnisse psychologischer Messungen nicht vollständig stabil sind, sondern Zufallsschwankungen unterliegen. Selbst wenn man sich vorstellt, man könnte dieselbe Person v mit demselben Test i unter denselben Bedingungen immer wieder testen, so setzt man nicht voraus, daß sie jedesmal den genau gleichen Testwert erzielt, sondern man nimmt an, daß der *beobachtete Testwert* X_{vi} mehr oder weniger stark schwankt. Der *wahre Wert* der Person v im Test i wird dann als derjenige Wert definiert, den die Person bei gedachter unendlicher Testwiederholung im Durchschnitt erreichen würde. Er wird mit τ (griechisch: tau) bezeichnet und ist der Erwartungswert zum beobachteten Testwert x_{vi}:

[2.1] $$\tau_{vi} = E(X_{vi})$$

Mit der Definition des wahren Werts als Erwartungswert bei gedachter unendlicher Meßwiederholung unter denselben Bedingungen, wurde dieser Begriff von viel unnötigem inhaltlichen Ballast befreit, der zunächst damit verbunden zu sein scheint. Der Ausdruck “wahrer Wert” legt vom Wortlaut her die Annahme nahe, es handle sich um einen idealen Wert, der der Person unabhängig vom Meßinstrument “in Wahrheit” zukommt, und der die Person zeitlich unveränderlich kennzeichnet. Wenngleich solche Vorstellungen in der älteren Literatur eine Rolle gespielt haben (Näheres dazu findet man bei Lord & Novick, 1968, Kapitel 2.9), so sind sie in der in [2.1] gegebenen Definition nicht mehr enthalten. Der wahre Wert ist für den speziellen Test spezifisch: Werden z.B. durch Hinzufügen von Wahlalternativen die Ratemöglichkeiten reduziert, so ändert sich dadurch (außer bei Personen, die nie raten) die zu erwartende Trefferzahl, also der wahre Wert im Sinn der Definition [2.1]. Wenn die Person nicht unter denselben, sondern unter anderen Bedingungen (nach Lern- oder Reifungsprozessen, bei geänderter Motivationslage usw.) getestet wird, kann sie einen anderen wahren Wert haben.

Da nun aber praktisch eine Testwiederholung unter genau gleichen Bedingungen nicht möglich ist, schon gar nicht unendlich oft, bleibt der wahre Wert eine theoretische Größe, die zwar geschätzt (zur Berechnung des Konfidenzintervalls zur Schätzung des wahren Werts siehe Formel [2.9] in diesem Kapitel), aber nie genau angegeben werden kann. Als *Meßfehler* wird die Differenz zwischen dem beobachteten Testwert der Person und dem theoretisch definierten wahren Wert bezeichnet:

[2.2] $$F_{vi} = X_{vi} - \tau_{vi}$$

Aus dieser Definition ergibt sich, daß für jeden Probanden der Erwartungswert des Meßfehlers Null ist:

[2.3] $$E(F_{vi}) = E(X_{vi} - \tau_{vi}) = E(X_{vi}) - \tau_{vi} = \tau_{vi} - \tau_{vi} = 0$$

In einer Population von Probanden, z.B. einem Altersjahrgang, werden sich die Personen in ihren wahren Werten T unterscheiden. Daraus, daß aus dem wahren Wert einer Person der Meßfehler nicht vorhersagbar ist (für jede Person - gleichgültig welchen wahren Wert sie hat - ist der Erwartungswert der Meßfehler Null) ergibt sich als weitere Folgerung, daß in jeder beliebigen Population die Meßfehler mit den wahren Werten des Tests unkorreliert sind. Darüber hinaus wird angenommen, daß die Meßfehler eines Tests i auch nicht mit den wahren Werten oder den Meßfehlern eines anderen Tests j korrelieren. Sie sind vielmehr als unsystematische Zufallsschwankungen aufzufassen. Diese grundlegenden Annahmen der klassischen Testtheorie sind in vier Axiomen zusammengefaßt:

Axiom I: Der Erwartungswert des Meßfehlers ist Null.
$E(F_i) = 0$.

Axiom II: Die Meßfehler korrelieren nicht mit den wahren Werten in demselben Test.
$\rho(F_i,T_i) = 0$.
(ρ = griechisch: rho)

Axiom III: Die Meßfehler im Test i korrelieren nicht mit den Meßfehlern in einem anderen Test j.
$\rho(F_i, F_j) = 0$.

Axiom IV: Die Meßfehler im Test i korrelieren nicht mit den wahren Werten aus einem anderen Test j.
$\rho(F_i,T_j) = 0$.

Die Axiome sind Ausgangspunkt für alle weiteren mathematischen Ableitungen. Wenn man Formeln aus der klassischen Testtheorie benutzt, hat man daher zu überlegen, ob die in den Axiomen ausgesprochenen Grundannahmen im vorliegenden Anwendungsfall zutreffen. Die Annahmen über die Unabhängigkeit der Meßfehler mögen zwar in der Regel plausibel sein, doch können in Spezialfällen durch mathematische Abhängigkeiten zwischen den Skalen auch die Meßfehler abhängig werden (z.B. durch Verrechnen derselben Items auf mehreren Skalen; weitere Beispiele findet man bei Stelzl, 1982, Kapitel 5.2).

2.2 Die Gütekriterien der klassischen Testtheorie: Objektivität, Reliabilität, Validität

Aufbauend auf den Begriffen “beobachteter Wert”, “wahrer Wert” und “Meßfehler” lassen sich Reliabilität und Validität, die beiden zentralen Gütekriterien der klassischen Testtheorie, definieren. Inhaltliche Voraussetzung für Reliabilität und Validität ist die Objektivität. Sie soll deshalb vorab behandelt werden.

2.2.1 Objektivität

Ein Testergebnis ist *objektiv*, wenn es nicht vom Testleiter (seiner Person, seinem Verhalten bei der Durchführung oder seinem Ermessen bei der Auswertung) abhängt. Der Test soll ja Aussagen über den Probanden machen, und nicht über den Psychologen, der ihn anwendet. Lienert (1961) unterscheidet Durchführungs-, Auswertungs- und Interpretationsobjektivität:

Durchführungsobjektivität bedeutet, daß das Testergebnis nicht davon abhängt, wer als Untersucher den Test mit dem Probanden durchführt. Um das zu erreichen, werden die Instruktionen zumindest sinngemäß, meist sogar wörtlich festgelegt, werden Abbruchzeiten bei nicht erfolgten Antworten festgesetzt, werden zulässige Hilfen und Kommentare möglichst im Wortlaut fixiert, usw. Durchführungsobjektivität ist verständlicherweise leichter zu erreichen, wenn der Proband nach der Instruktion relativ selbständig weiterzuarbeiten hat, als wenn die Durchführung in ständiger Interaktion mit dem Versuchsleiter erfolgt, wie das z.B. bei Tests mit jüngeren Kindern erforderlich ist.

Die Kontrolle der Durchführungsobjektivität ist vom Aufwand der Datenerhebung her gesehen relativ schwierig: Der theoretisch gesehen einfachste Weg, nämlich denselben Test an denselben Probanden mehrmals mit wechselnden Versuchsleitern durchzuführen, führt nicht nur zu einer erheblichen zeitlichen Belastung des Probanden, sondern kommt meist auch wegen massiver Erinnerungs- und Wiederholungseinflüsse kaum in Betracht. Ein anderer Weg besteht darin, die Probanden den Untersuchern zufällig zuzuordnen und nach Mittelwertsunterschieden zwischen den Untersuchern zu fragen. Wenn der Test hohe Durchführungsobjektivität hat, sollten keine Mittelwertsunterschiede auftreten. Allerdings dürfte auch eine Zufallszuordnung von Probanden zu Untersuchern meist erhebliche organisatorische Probleme mit sich bringen. Wegen solcher praktischer und versuchstechnischer Schwierigkeiten wird Durchführungsobjektivität auch weit seltener untersucht als Auswertungsobjektivität.

Auswertungsobjektivität ist gegeben, wenn bei vorliegendem Testprotokoll (Antworten des Probanden) das Testergebnis (IQ, Punktwert o.ä.) nicht von der Person des Testauswerters abhängt. Bei Tests mit Mehrfachwahl-Aufgaben ist Auswertungsobjektivität im allgemeinen problemlos zu erreichen. Wenn dagegen der Proband die Antwort selbst zu formulieren hat und der Spielraum möglicher Antworten groß ist, müssen detaillierte Auswertungsregeln erarbeitet werden, und die Auswertungsobjektivität muß empirisch überprüft werden. Aber auch dann, wenn der Proband relativ komplexe Probleme zu bearbeiten hat, und entsprechend unterschiedliche Teillösungen möglich sind, kann hohe Auswertungsobjektivität erreicht werden. Versuchspläne zur Bestimmung der Auswertungsobjektivität findet man bei Nußbaum (1987).

Interpretationsobjektivität liegt vor, wenn verschiedene Psychologen aufgrund desselben Testwertes zu denselben Schlußfolgerungen kommen. Hier kann der Testautor zwar Hilfestellungen geben, indem er z.B. möglichst umfangreiche Angaben zur Validität macht und durch ausführliche Testnormen den Vergleich des Probanden mit einschlägigen Bezugsgruppen ermöglicht - bei der Vielzahl möglicher Fragestellungen und möglicher Rahmenbedingungen wird eine vollständige Interpretationsobjektivität aber kaum zu erreichen sein.

2.2.2 Reliabilität

Die Reliabilität ist eines der zentralen Gütekriterien der klassischen Testtheorie. Es geht dabei um die Meßgenauigkeit im Sinne der Reproduzierbarkeit des Testergebnisses bei konstanten Bedingungen. Die Frage, was gemessen wird, ob z.B. ein Intelligenztest wirklich Intelligenz mißt oder nur Schulwissen, bleibt dabei noch ausgeklammert. Wenn ein Test hohe Meßgenauigkeit haben soll, dürfen Zufallseinflüsse (Meßfehler im Sinn der Axiome) nur eine geringe Rolle spielen. Aus den Axiomen läßt sich ableiten, daß sich die beobachtete Testvarianz aus der Varianz der wahren Werte und der Varianz der Meßfehler zusammensetzt:

[2.4] $$\sigma^2(X) = \sigma^2(T) + \sigma^2(F)$$

Die Reliabilität ist definiert als:

[2.5] $$Rel = \frac{\sigma^2(T)}{\sigma^2(X)}$$

Sie gibt an, welcher Teil der beobachtbaren Testvarianz auf die Varianz der wahren Werte zurückzuführen ist.

Wenn man an einer Personenpopulation den Test zweimal unter denselben Bedingungen durchführt, so daß für jede Person zwei parallele Messungen X und X' vorliegen, so ist die Reliabilität gleich der Korrelation der beiden parallelen Messungen:

[2.6] $$Rel = \rho(X,X')$$

In der Anwendung steht man allerdings vor dem Problem, daß eine Meßwiederholung unter genau denselben Bedingungen nicht durchführbar ist, da als Folge der ersten Testdurchführung Erinnerungs-, Übungs-, Ermüdungseinflüsse usw. auftreten. Als eine näherungsweise Realisierung kommt eine Wiederholung desselben Tests nach einem längeren oder kürzeren Zeitintervall *(Testwiederholungsreliabilität)* in Betracht, oder auch die Vorgabe von zwei verschiedenen, nach bestem Wissen als parallel konzipierten und auf Parallelität geprüften Testformen *(Paralleltestreliabilität)*. Wenn nur eine Testvorgabe vorliegt, kann man auch durch geeignete Unterteilung dieses einen Tests Aufschluß über die Reliabilität erhalten (Berechnung der *Testhalbierungsreliabilität* bei Teilung des Tests in zwei Teile, der *inneren Konsistenz* bei Teilung in mehr als zwei Teile).

Verschiedene Arten der Reliabilitätsbestimmung stimmen in ihren Ergebnissen meist nicht genau überein. Man kann das als eine Unzulänglichkeit der Anwendung beklagen, die eine Wiederholung unter gleichen Bedingungen eben nur annäherungsweise ermöglicht. Man kann aber auch aus dem Vergleich der auf verschiedene Arten erhobenen Reliabilitätskoeffizienten wichtige Informationen über den Test gewinnen: Die Korrelation der Testergebnisse bei Testwiederholung nach unterschiedlich langem Zeitabstand ist in jedem Fall von Interesse, ebenso die Korrelation zwischen als gleichwertig angebotenen Parallelformen eines Tests. Testwiederholungsreliabilitäten mit unterschiedlich langen Zeitintervallen zwischen erster und zweiter Testdurchführung sagen auch etwas über die Stabilität des gemessenen Merkmals aus. Wenn das Zeitintervall kurz ist, muß man bei der vergleichenden Interpretation der Koeffizienten auch an Erinnerungs- und Übungseinflüsse denken. Dabei ist zu berücksichtigen, daß Einflußgrößen, die sich bei allen Probanden gleich auswirken (z.B. zu einem für alle Personen gleichen Zuwachs von 5 Punkten führen), nur zu einer Mittelwertsver-

schiebung führen, sich aber auf die Korrelation nicht auswirken. Nur Einflußgrößen mit individuell unterschiedlicher Wirkung (z.B. ein individuell unterschiedlicher Lernzuwachs, ein individuell unterschiedlicher Vorteil durch Erinnerung an bereits gefundene Lösungen) beeinflussen die Korrelation.

Die Störquellen bei der Bestimmung der Paralleltestreliabilität sind im wesentlichen die gleichen. Da nicht genau derselbe Test ein zweites Mal vorgegeben wird, werden Übungs- und Erinnerungseinflüsse sich nicht ganz so stark auswirken, dafür kommt mangelnde Parallelität der als parallel konzipierten Tests als mögliche Störquelle hinzu.

Die Testhalbierung wird gerne aus Gründen der Ökonomie angewendet, weil nur eine Testvorgabe erforderlich ist. Die Items werden auf zwei möglichst parallele Testhälften verteilt, jede Testhälfte für sich ausgewertet und die Korrelation der Testwerte aus den beiden Hälften bestimmt. Damit hat man eine Schätzung der Reliabilität des Tests bei halber Länge. Daraus wird mit Hilfe der Spearman-Brown-Formel (siehe Fischer, 1974, S.50) die Reliabilität für den ganzen Test errechnet. Zu beachten ist, daß situative Effekte (Tagesverfassung des Probanden, äußere Umstände der Testdurchführung) beide Testhälften in gleicher Weise betreffen und damit zur Korrelation beitragen können. Diese Effekte gehen hier in die Varianz der wahren Werte, nicht in die Varianz der Meßfehler ein. Testhalbierungskoeffizienten fallen deshalb meist höher aus als Testwiederholungs- oder Paralleltestkoeffizienten mit einem Zeitintervall zwischen den beiden Testdurchführungen.

Eine beliebte Art der Testhalbierung ist die Odd-Even-Methode. Dabei werden die Items durchnummeriert und dann die Items mit ungeradzahliger Nummer (englisch: odds) in die eine Testhälfte, die mit geradzahliger (englisch: even) Nummer in die andere Testhälfte gerechnet. Dabei betreffen dann auch individuelle Schwankungen im Leistungsverlauf (z.B. Anfangshemmung, Übung, Ermüdung, zwischenzeitliche Schwankungen in Motivation und Aufmerksamkeit) beide Testhälften in gleicher Weise und können zur Erhöhung der Korrelation beitragen. Wird die Odd-Even-Methode auf reine Geschwindigkeitstests (die Personen unterscheiden sich nur darin, wie weit sie in der vorgegebenen Zeit mit der Bearbeitung gekommen sind; Fehler kommen kaum vor) angewendet, so stimmt die Zahl der Richtigen in den beiden Testhälften trivialerweise fast genau überein (bei ungerader Zahl von Bearbeiteten gibt es bei den Ungeradzahligen um ein richtiges Item mehr, sonst stimmen die Testhälften exakt überein) und die Korrelation wird (fast) Eins sein. Diese Korrelation besagt aber nichts über die Reproduzierbarkeit des Testergebnisses bei gedachter Wiederholung unter denselben Bedingungen, ist also als Reliabilitätsschätzung ungeeignet. Als Alternative kommt eine Halbierung nach der Testzeit (nach der Hälfte der Zeit wird ein Signal gegeben und die Probanden kennzeichnen mit einem Strich, wie weit sie bis dahin gekommen sind) in Betracht oder eine der oben genannten anderen Arten der Reliabilitätsbestimmung (Wiederholung, Paralleltest). Auch bei Tests, die zwar keine reinen Geschwindigkeitstests sind, bei denen der Zeitdruck aber doch eine erhebliche Rolle spielt, sollte die Odd-Even-Methode nicht verwendet werden, oder zumindest durch andere Arten der Reliabilitätsbestimmung ergänzt werden.

Die innere Konsistenz hängt im wesentlichen von den Korrelationen der Testteile (Items) untereinander ab und gibt somit Auskunft darüber, inwieweit der Test in sich homogen ist. Letzteres ist oft auch zum Zusammenhang mit der Validität (der Frage, was der Test mißt) von Interesse. Wenn die Items eines Tests unkorreliert sind, der Test also extrem heterogen ist, so ist die innere Konsistenz Null. Das gilt auch dann,

wenn jedes Item für sich genommen perfekt reliabel ist, der Test also keine Fehlervarianz enthält und eine Reliabilität von Eins hat. Die innere Konsistenz unterschätzt in einem solchen Fall die Reliabilität. In der Praxis hat man es allerdings meist mit relativ homogenen Tests zu tun (man will Ähnliches in einen Testwert zusammenfassen und nicht Äpfel und Birnen addieren). Zudem stammen die Daten aus nur einer Testdurchführung, so daß (ähnlich wie bei der Testhalbierung) situative Bedingungen alle Testteile in gleicher Weise betreffen. Erfahrungsgemäß fallen Reliabilitätsschätzungen mittels interner Konsistenz verglichen mit Testwiederholungskoeffizienten nach einem längeren Zeitintervall meist höher aus.

Übersicht 2.1: Mögliche Störeinflüsse bei verschiedenen Arten der Reliabilitätsbestimmung

Testwiederholungsmethode: Erinnerung, Übung, bei längerem Zeitintervall auch Veränderungen des Merkmals.
Paralleltestmethode: Störeinflüsse wie bei Testwiederholung, zusätzlich: mangelnde Parallelität der als Parallelformen angebotenen Tests.
Testhalbierungsmethode: Situative Einflüsse (z.B. Tagesverfassung der Person) betreffen beide Testhälften in gleicher Weise und erhöhen die Korrelation. Bei einer Halbierung nach der Odd-Even-Methode gilt das auch für individuelle Leistungsschwankungen im Verlauf der Versuchsdurchführung. Bei Speed-Tests führt die Odd-Even-Methode zu einer Überschätzung der Reliabilität und sollte nicht angewandt werden.
Innere Konsistenz (Teilung in beliebig viele Teile; meist: Teilung in Einzelaufgaben): Situative Einflüsse (Tagesverfassung) betreffen alle Teile (Items) und können die Korrelationen der Teile (Iteminterkorrelationen) und damit die innere Konsistenz erhöhen. Heterogene Tests (Tests, bei denen die wahren Werte der Testteile niedrig korrelieren, weil jeder Teil etwas anderes mißt) ergeben auch bei hoher Reliabilität niedrige Werte für die innere Konsistenz. Ihre Reliabilität wird unterschätzt.

Hat man nun für eine Personenpopulation die Testvarianz und die auf die eine oder andere Art bestimmte Reliabilität vorliegen, so kann man aus diesen beiden Angaben die Fehlervarianz berechnen (die Ableitung von Formel [2.7] ergibt sich aus [2.4] und [2.5]):

$$\sigma^2(F) = \sigma^2(X) \cdot (1 - Rel) \qquad [2.7]$$

Die Wurzel aus der Fehlervarianz heißt *Standardmeßfehler*.

$$\sigma(F) = \sigma(X) \cdot \sqrt{(1-Rel)} \qquad [2.8]$$

Nimmt man weiter an, daß die Meßfehler normalverteilt sind und daß die Fehlervarianz in allen Skalenbereichen gleich groß ist, so kann man mithilfe des Standardmeßfehlers ein Konfidenzintervall für den wahren Wert eines Probanden angeben:

Ausgehend von der Überlegung, daß in einer Normalverteilung 95% der Fälle in einem Bereich ± 1.96 Streuungseinheiten liegen, kann man zunächst feststellen, daß ein Proband mit einem wahren Wert z_{vi} mit 95%iger Sicherheit einen beobachteten Wert im Bereich

$$\tau_{vi} \pm 1.96\,\sigma(F)$$

erzielt.

Daraus läßt sich ableiten, daß die Grenzen des Konfidenzintervalls

[2.9] $$X_{vi} - 1.96\,\sigma(F) \leq \tau_{vi} \leq X_{vi} + 1.96\,\sigma(F)$$

mit 95%iger Sicherheit den wahren Wert eines Probanden einschließen.

Auf ähnlichen Überlegungen aufbauend kann man auch für komplexere Maße Konfidenzintervalle ableiten. Man kann z.B. eine kritische Differenz berechnen, die überschritten werden muß, damit der Unterschied zwischen zwei beobachteten Testwerten (bei der gewählten Irrtumswahrscheinlichkeit α) nicht mehr als meßfehlerbedingt anzusehen ist (siehe Kapitel 3.2). Die Gültigkeit dieser Formeln setzt-wie oben erwähnt - gleiche Fehlervarianz in allen Skalenbereichen voraus. Diese Voraussetzung, die auch als Homoskedastizitäts-Annahme bezeichnet wird, ist empirisch prüfbar (indem man z.B. die Fehlervarianzen aus unterschiedlichen Teilpopulationen von Probanden berechnet) und praktisch von großer Relevanz. Trotzdem wird ihr bei der Testkonstruktion erstaunlich wenig Aufmerksamkeit geschenkt: Kaum ein Testmanual enthält hierzu explizite Angaben.

2.2.3 Validität

Hier geht es um die Frage, ob der Test das erfaßt, was er erfassen soll. Formal könnte man die Validität als Korrelation der Testwerte mit der Eigenschaft, die gemessen werden soll, definieren. Praktisch wird man die Validität allerdings nicht mit einer einzigen Korrelation ausdrücken können, sondern je nach Testinhalt und Anwendungsbereich eine Fülle von Angaben zusammentragen müssen, die in ihrer Gesamtheit darüber Aufschluß geben, inwieweit der Test mißt, was er messen soll.

Am einfachsten scheint die Frage nach der Validität dann beantwortet zu sein, wenn die Testaufgaben selbst eine Stichprobe aus dem Verhaltensbereich sind, über den eine Aussage getroffen werden soll: z.B., wenn Rechtschreibkenntnisse durch ein Diktat abgeprüft werden. In solchen Fällen spricht man von *inhaltlicher Validität* (content validity), bisweilen auch von *logischer Validität.* Inhaltliche Validität sollte freilich nicht nur aufgrund des Augenscheins (zur sog. Augenscheinvalidität siehe unten) beansprucht werden: In einem Diktat könnte z.B. der gewählte Text nicht repräsentativ sein, weil viele Fremdwörter aus einem engen Spezialgebiet vorkommen oder weil bestimmte Rechtschreibregeln nicht zur Anwendung kommen. Die Frage, wie inhaltliche Validität zu erreichen ist, wurde speziell im Zusammenhang mit lehrzielorientierten Tests viel diskutiert. Eine zusammenfassende Darstellung findet man bei Klauer (1987).

Mit inhaltlicher Validität leicht zu verwechseln ist die *Augenscheinvalidität* (face validity), zumal inhaltlich validen Tests in aller Regel auch Augenscheinvalidität zukommt. Augenscheinvalidität gibt an, inwieweit der Validitätsanspruch eines Tests einem Laien “vom bloßen Augenschein her” gerechtfertigt erscheint. Ein Intelligenztest z.B. hat hohe Augenscheinvalidität, wenn der Laie es aufgrund von Inhalt und Gestaltung des Tests für plausibel hält, daß damit Intelligenz gemessen werden kann. Unter wissenschaftlichem Gesichtspunkt mag Augenscheinvalidität zunächst als gänzlich irrelevant erscheinen. Es ist jedoch zu bedenken, daß für die Mitteilbarkeit

Beispiel 2.1: Verschiedene Arten der Validitätsbestimmung: Angaben aus der Handanweisung zum Zahlen-Verbindungs-Test nach Oswald & Roth (1978).

Der Zahlen-Verbindungs-Test (ZVT) nach Oswald & Roth (1978) wurde mit dem Anspruch entwickelt, durch Messung der kognitiven Leistungs- und Verarbeitungsgeschwindigkeit ein - wenn auch spezifischer - Intelligenztest zu sein.
Die Aufgabe des Probanden besteht darin, auf einem Blatt, das mit Zahlen bedruckt ist, die Ziffern der Reihe nach (1,2,3... usw.) aufzufinden und miteinander zu verbinden. Es sind insgesamt vier Blätter mit möglichst hohem Arbeitstempo zu bearbeiten. Gemessen wird die benötigte Zeit (oder bei vorgegebener Zeit die Anzahl der verbundenen Zahlen).
Vom Inhalt der Aufgabenstellung her ist nicht ohne weiteres ersichtlich, daß es sich um einen Intelligenztest handelt. Man könnte ebensogut an einen Konzentrationstest denken. Der Test kann also als Intelligenztest nur begrenzt Augenscheinvalidität beanspruchen. Als empirische Belege für die Validität als Intelligenztest sind u.a. Korrelationen mit fünf verschiedenen, bekannten Intelligenztests an unterschiedlichen Stichproben erhoben worden (konvergente Validität). Die Korrelationen fallen je nach Stichprobe mittel bis hoch aus. Für zwei Tests, nämlich PSB (=Prüfsystem für Schul- und Bildungsberatung nach Horn, 1969) und IST (=Intelligenz-Struktur-Test nach Amthauer, 1970) liegen auch Angaben für Stichproben vor, die jeweils für einen Altersjahrgang repräsentativ zusammengesetzt wurden. Dort liegen die Korrelationen zum ZVT um 0,7 bis 0,8.
Weiter soll belegt werden, daß der ZVT nicht im wesentlichen nur Konzentrationsfähigkeit oder nur Handgeschwindigkeit erfaßt. Dazu wurden, wieder an unterschiedlichen Stichproben, Korrelationen zu bekannten Konzentrationstests und zu einem Test der Handgeschwindigkeit (Striche-Ziehen als reine Geschwindigkeitsaufgabe) berechnet. Die Korrelationen zu Konzentrationstests fallen deutlich niedriger aus als zu Intelligenztests, die Korrelationen zum Striche-Ziehen schwanken um Null (diskriminante Validität). Als weiterer Beitrag zur Konstruktvalidität wurde der ZVT mit den Untertests aus verschiedenen bekannten Intelligenztests zusammen einer Faktorenanalyse unterzogen. Der ZVT zeigte die höchsten Ladungen in einem Faktor, der als "Kognitive Leistungsfähigkeit" interpretiert wurde (zu Grundgedanken der Faktoranalyse, Begriff der Ladung und Vorgehen bei der Interpretation siehe Kapitel 4.2).
Die bisher genannten Methoden (Berechnung von Korrelationen als Angaben zur konvergenten und diskriminanten Validität, Faktoranalysen) sind Standardmethoden, um Konstruktvalidität zu belegen. Darüber hinaus enthält die Handanweisung eine Reihe von weiteren Angaben (über Beziehungen zu EEG-Variablen, über Übungseffekte, über Stadt-Land-Unterschiede, Unterschiede zwischen Heimkindern und anderen Hauptschülern, usw.), die zwar jede für sich noch keinen Validitätsbeleg darstellen, die aber in ihrer Gesamtheit doch mit dazu beitragen, abzugrenzen, was der Test mißt.
Angaben zur prognostischen Validität sind in der Handanweisung kaum zu finden. Es werden einige Korrelationen zu Schulnoten und zu Schulleistungstests mitgeteilt, die relativ niedrig ausfallen. Über den zeitlichen Abstand zwischen der Durchführung des ZVT und der Erhebung der Schulleistung ist nichts gesagt, so daß zu vermuten ist, daß es sich um eine gleichzeitige Erhebung handelt und nicht über eine Prognose im engeren Sinn (Vorhersage späterer Leistungen).

des Testergebnisses und für die Akzeptanz von Seiten des Probanden der Augenscheinvalidität eine ganz erhebliche Bedeutung zukommen dürfte.

Die meisten psychologischen Tests zielen darauf ab, relativ komplexe psychologische Konstrukte (Fähigkeiten, Einstellungen, Persönlichkeitsmerkmale) zu erfassen, deren Bedeutung im Rahmen einer mehr oder weniger detailliert ausgearbeiteten psychologischen Theorie beschrieben wird. Die Tests sind hier in der Regel Indikatoren und nicht einfach Verhaltensstichproben, so daß inhaltliche Validität nicht in Anspruch genommen werden kann. Um zu zeigen, daß der Test das angepeilte Konstrukt erfaßt, also *Konstruktvalidität* (construct validity) besitzt, können Belege verschiedener Art herangezogen werden: Die Testergebnisse können mit anderen Indikatoren für dasselbe Konstrukt (Tests mitähnlichem Geltungsanspruch; Beurteilungen durch Klassenkameraden, Eltern, Lehrer; Verhaltensbeobachtung in einschlägigen Situationen) korreliert werden. Fallen diese Korrelationen hoch aus, so spricht man von konvergenter Validität oder *Übereinstimmungsvalidität* (convergent validity). Darüber hinaus kann man untersuchen, ob sich das Konstrukt von bedeutungsähnlichen Konstrukten hinreichend abgrenzen und im Test hinlänglich frei von unerwünschten Komponenten erfassen läßt. So z.B. könnte man fragen, ob sich Kreativität begrifflich von allgemeiner Intelligenz hinreichend klar abgrenzen läßt, und ob ein bestimmter Test nicht an Stelle von Einfallsreichtum zu einem hohen Teil Schreibgeschwindigkeit erfaßt. Solche Fragen der Abgrenzung gegen bedeutungsverwandte Konstrukte und gegen irrelevante Komponenten im Test sind Fragen nach der *diskriminanten Validität* (discriminant validity). Konvergente und diskriminante Validität werden häufig mithilfe von Faktorenanalysen untersucht, wobei sich in Verbindung mit entsprechenden Datenerhebungsplänen konfirmatorische Faktoranalysen anbieten (Näheres siehe Kapitel 4.2.3). Beiträge zur Konstruktvalidierung können aber auch auf ganz anderen Wegen geleistet werden: Auch Effekte experimenteller Variation (Beeinflussung der Motivation durch zusätzliche Anreize, der Lösungsstrategie durch spezielle Instruktion, Einführung/Aufhebung von Zeitdruck usw.) können mit darüber Aufschluß geben, was der Test mißt.

Über die Frage nach der Konstruktvalidität hinausgehend stellt sich für den Praktiker die Frage, mit welchem Erfolg sich der Test in der diagnostischen Praxis einsetzen läßt. Ihn interessiert, wie die Testprognose mit später anfallenden Bewährungskriterien korreliert, also die *prognostische Validität.* Sind Test- und Kriteriumswert bivariat normalverteilt, so lassen sich die Kriteriumswerte mittels linearer Regression aus den Testwerten vorhersagen:

$$[2.10] \quad Y^* = \beta X + \alpha$$

wobei : Y^* = vorhergesagter Kriteriumswert
X = Testwert

$\beta = \rho_{xy} \frac{\sigma_y}{\sigma_x}$ = Regressionskoeffizient

ρ_{xy} = Korrelation zwischen Test und Kriterium

$\alpha = \mu_y - \beta\mu_x$ = Regressionskonstante

μ_x , μ_y = Mittelwerte von Test und Kriterium

(α = griechisch: alpha, β = griechisch: beta, μ = griechisch: my)

Die Genauigkeit der Schätzung läßt sich mithilfe des *Standardschätzfehlers* $o_{y/x}$ angeben:

[2.11] $$\sigma_{y/x} = \sigma_y \sqrt{(1 - \rho_{xy}^2)}$$

Mit einer Sicherheit von 0.95 liegt der Kriteriumswert in einem Bereich von

$$Y^* \pm 1.96\, \sigma_{y/x}$$

Diese Genauigkeitsangabe setzt voraus, daß die Streuung um die Regressionslinie überall gleich ist und daß die Abweichungen vom Regressionsschätzwert jeweils normalverteilt sind. *(Homoskedastizität* der Regression von Y auf X). Diese Voraussetzungen sind erfüllt, wenn Test und Kriterium - wie oben angenommen - bivariat normalverteilt sind. Sie wären nicht erfüllt, wenn z.B. im unteren und mittleren Bereich ein enger Zusammenhang zwischen Test- und Kriteriumswerten bestünde, nicht aber im oberen Bereich, und mithin die Vorhersagegenauigkeit in den einzelnen Bereichen stark unterschiedlich wäre. Auch bei Tests mit eng umschriebenem Einsatzbereich, wie z.B. Schuleingangstests, sind viele unterschiedliche Bewährungskriterien erhebbar: Man kann den Schulerfolg nach unterschiedlichen Zeitintervallen erheben, man kann Noten, Lehrerurteil oder Schulleistungstests heranziehen, man kann Elternauskünfte über Schulangst oder Schulunlust einholen, usw. Auch hier ist also die Validität nicht auf die Angabe einer einzigen Zahl beschränkt. Bei Tests, wie z.B. allgemeinen Intelligenztests, die im Zusammenhang mit recht unterschiedlichen Fragestellungen zum Einsatz kommen können, ist die Zahl möglicher Kriterien für prognostische Validität unbegrenzt, und die Aufgabe, prognostische Validität zu untersuchen, grundsätzlich nicht abschließbar.

2.2.4 Beziehungen zwischen Reliabilität und Validität

Wenn die Validität als Korrelation zwischen einem Test X (z.B. einem Konzentrationstest) und einem Kriterium Y (z.B. der Schulleistung, erfaßt mit einem bestimmten Schulleistungstest) bestimmt wird, so hängt diese Korrelation nicht nur davon ab, wie eng das, was der Test mißt (die Konzentrationsfähigkeit), mit dem, was das Kriteriumsmaß erfaßt (der Schulleistung), zusammenhängt, sondern auch von der Reliabilität der beiden Maße. Selbst wenn die Konzentrationsfähigkeit eine wesentliche Grundlage der Schulleistung darstellt, kann der beobachtete Zusammenhang nicht hoch ausfallen, wenn beide Maße einen hohen Anteil an zufälligen Meßfehlern enthalten, also unreliabel sind.

Es läßt sich zeigen (siehe Lord & Novick, 1968, Kapitel 3.9), daß zwischen der Korrelation der wahren Werte T_x und T_y und der Korrelation der beobachteten Werte X und Y folgende Beziehung besteht:

[2.12] $$\rho\,(T_x T_y) = \frac{\rho(X\,Y)}{\sqrt{\mathrm{Rel}(X)\,\mathrm{Rel}(Y)}}$$

Für die Korrelation von T_x mit Y bzw. von T_y mit X gilt

[2.13a] $$\rho(T_x Y) = \frac{\rho(X\,Y)}{\sqrt{Rel(X)}}$$

[2.13b] $$\rho(T_y X) = \frac{\rho(X\,Y)}{\sqrt{Rel(Y)}}$$

Formel [2.12] wird als doppelte, Formel [2.13a] bzw. [2.13b] als einfache Minderungskorrektur bezeichnet, bisweilen auch als *Verdünnungsformel* (als wörtliche Übersetzung des englischen Ausdrucks *Correction for Attenuation).* Die minderungskorrigierten Korrelationen werden bisweilen auch mit dem mathematischen Symbol für "unendlich" als $\rho_{\infty\infty}$ (doppelte Minderungskorrektur nach Formel [2.12] und $\rho_{\infty y}$ (einfache Minderungskorrektur nach Formel [2.13a]) geschrieben. Diese Schreibweise nimmt darauf Bezug, daß perfekte Reliabilität (der Test - analog: das Kriterium - besteht nur aus wahren Werten) theoretisch durch unendliche Testverlängerung erreicht werden könnte (zum Zusammenhang zwischen Testlänge und Reliabilität siehe Lord & Novick, 1968, Kapitel 5.10; Fischer 1974, Kapitel 4).

Mithilfe der Minderungskorrektur ist es also möglich, Korrelationen zwischen wahren Werten zu berechnen, obwohl man für keine einzige Person den wahren Wert kennt. Solche minderungskorrigierte Korrelationen interessieren
(a) in der Grundlagenforschung: Wenn man sich für die Determinanten der Schulleistung interessiert, so will man von der Frage der Reliabilität der speziellen Test- und Kriterienmaße absehen und wird deshalb die Korrelation der wahren Werte als Korrelation der beiden Fähigkeiten berechnen.
(b) bei der Testkonstruktion: Formel [2.13a] gibt darüber Auskunft, inwieweit durch Reliabilitätsverbesserung (z.B. durch Verlängern des Tests durch Hinzufügen weiterer Aufgaben, Ausschalten von Ratemöglichkeiten, genauere Festlegung von Auswertungsregeln usw.) die prognostische Validität des Tests gesteigert werden kann. Im praktisch unrealistischen Idealfall, wenn es gelänge, die Reliabilität auf den Wert Eins zu steigern, würde die Validität auf den in Formel [2.13a] errechneten Wert steigen. Ist dieser Wert zu niedrig, so kann eine weitere Verbesserung nur durch eine Änderung des Testinhalts oder Hinzufügen weiterer Tests (siehe Kapitel 4) erreicht werden, nicht aber durch bloße Reliabilitätsverbesserung am vorliegenden Test.
Analog dazu gibt Formel [2.13b] Auskunft, inwieweit die Validität maximal steigen könnte, wenn bei unverändert belassenem Test X das Kriteriumsmaß perfekt reliabel gemacht werden könnte. Ist der nach Formel [2.13b] errechnete Wert unbefriedigend, so kann dieses unbefriedigende Ergebnis nicht mehr auf bloße Reliabilitätsmängel bei der Erfassung des Kriteriums zurückgeführt werden. Die Revision muß an anderer Stelle (Reliabilitätsverbesserung des Tests, inhaltliche Verwandtschaft zwischen Test und Kriterium) ansetzen.

2.3 Zur Populationsabhängigkeit der klassischen Gütekriterien

Die Gütekriterien der klassischen Testtheorie, nämlich Objektivität, Reliabilität und Validität beziehen sich stets auf eine bestimmte Personenpopulation. Sie ändern sich, wenn die Population anders zusammengesetzt ist. Das ist am leichtesten am Beispiel

der Reliabilität zu erkennen: Die Reliabilität ist der Anteil der wahren Varianz an der Testvarianz (vgl. Kapitel 2.2):

[2.14] $$Rel = \frac{\sigma^2(T)}{\sigma^2(X)} = \frac{\sigma^2(T)}{\sigma^2(T)+\sigma^2(F)}$$

Setzt man voraus, daß die Fehlervarianz gleich bleibt, so ist die Reliabilität umso größer, je mehr wahre Varianz vorhanden ist. Wie man an Formel [2.14] sieht, ist die Reliabilität Null, wenn in einer extrem homogenen Population alle Personen denselben wahren Wert haben, so daß die wahre Varianz Null ist. Die beobachtete Testvarianz besteht dann nur aus Fehlervarianz. Ist dagegen die Population extrem heterogen, die Varianz der wahren Werte also sehr groß, so geht die Reliabilität gegen Eins. Mithilfe entsprechender Formeln ist es möglich, zu berechnen, wie sich die Reliabilität in Abhängigkeit von der in der Population vorhandenen Testvarianz ändert (Lord & Novick, 1968, Kapitel 6). Man muß dazu in einer Population Varianz und Reliabilität kennen (z.B. für einen bestimmten Altersjahrgang aus der Testhandanweisung entnehmen) und kann dann für eine andere Population (z.B. nur Oberschüler dieses Jahrgangs), von der man nur die Varianz kennt (aus der Testhandanweisung, aus eigenen Daten, oder nur als grobe Schätzung aufgrund von Erfahrungen mit ähnlichen Tests), berechnen, wie dort die Reliabilität ausfallen würde.

[2.15] $$Rel^* = 1 - \frac{\sigma_x^2}{\sigma_x^{2*}} \quad (1 - Rel)$$

Rel = bekannte Reliabilität in einer Population mit bekannter Testvarianz σ_x^2
Rel* = Reliabilität, die berechnet werden soll
σ_x^{2*} = Testvarianz in der Population, für die die Reliabilität (Rel*) berechnet werden soll

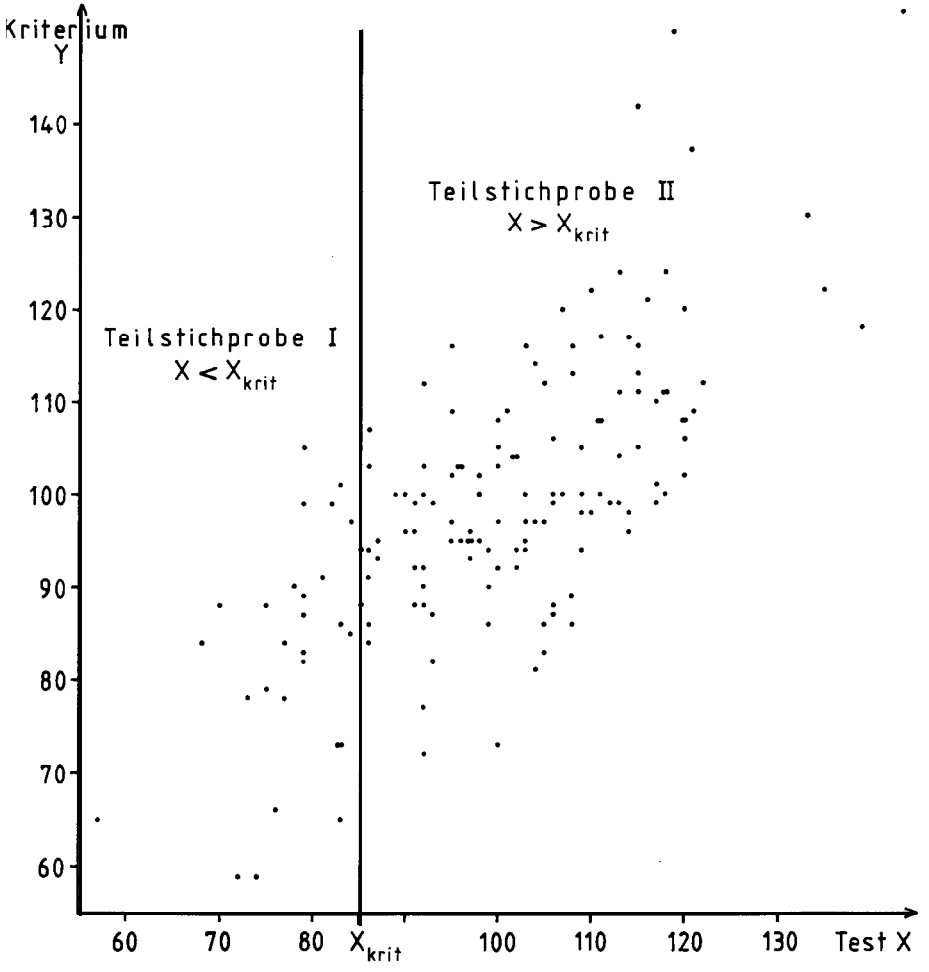
Ähnliches gilt für die Validität: Auch die Korrelation des Tests mit einem Validitätskriterium hängt von der Zusammensetzung der Personenpopulation ab. Unter bestimmten Voraussetzungen (Linearität der Regression des Kriteriums Y auf den Test X, gleiche Streuung der Kriteriumswerte um die Regression in allen Skalenbereichen) kann man berechnen, wie sich die Kriteriumskorrelation ändert, wenn sie an einer Population mit größerer oder kleinerer Testvarianz berechnet wird (Lord & Novick, 1968, Kapitel 6). Je größer die Testvarianz, desto höher fällt unter den genannten Voraussetzungen die Kriteriumskorrelation aus. Das ist in Abbildung 2.1 am Beispiel der bivariaten Normalverteilung zwischen Test und Kriterium veranschaulicht: Betrachtet man die gesamte Punktwolke, so ist die Korrelation zwischen Test und Kriterium r = 0.71. Betrachtet man nur die Probanden mit einem Testwert über X_{krit}, so fällt in dieser Teilpopulation die Korrelation niedriger aus. Das sieht man schon an der Form der Punktwolke, die eher kugelig und weniger langgestreckt ist, als die Punktwolke für die Gesamtpopulation. Berechnet man die Korrelation in der Teilpopulation, so erhält man r = .63.

Die bei Lord & Novick (1968, Kapitel 6.8) unter den genannten Voraussetzungen abgeleitete Formel über die Änderung der Validität in Abhängigkeit von der Testvarianz lautet:

[2.16] $$\rho^*_{xy} = \frac{1}{\sqrt{1 + \frac{\sigma_x^2}{\sigma_x^{2*}} \left(\frac{1}{\rho^2_{xy}} - 1\right)}}$$

ρ_{xy} = bekannte Validität in der Population mit Testvarianz σ_x^2
ρ^*_{xy} = zu berechnende Validität in der Population mit Testvarianz σ_x^{2*}

Abbildung 2.1: Korrelation und Selektion



Die Gesamtstichprobe wird nach einem kritischen Testwert X_{krit} geteilt. Während in der Gesamtstichprobe die Korrelation r = .71 beträgt, beträgt sie in Teilstichprobe I nur r_I = .42 und in Teilstichprobe II r_{II} = .63.

Die Objektivität kann in Form von Korrelationen zwischen Beurteilern oder auch in Anschluß an eine varianzanalytische Bestimmung in Form von Varianzkomponenten ausgedrückt werden. In jedem Fall hängt das Ergebnis von der Zusammensetzung der Probandenstichprobe ab, die es zu beurteilen galt. Unter sonst gleichen Umständen gilt: Je größer die Varianz in der Probandenstichprobe, desto höher die Kennwerte für die Objektivität.
Die Tatsache, daß die Gütekriterien der klassischen Testtheorie populationsbezogen definiert sind, wurde in den Siebzigerjahren von Vertretern des Latent- Trait-Ansatzes (z.B. Fischer, 1974, S. 137) als wichtiger Kritikpunkt ins Feld geführt, stellt aber unseres Erachtens keinen grundsätzlichen Mangel dar. Die Gütekriterien geben an, was der Test, angewendet auf eine bestimmte Population, zu leisten vermag. Es gibt, wie erwähnt, die Möglichkeit, die Koeffizienten auf Populationen mit anderer Varianz umzurechnen. Darüber hinaus gibt es auch im Rahmen der klassischen Testtheorie die Möglichkeit, Kennwerte zu benutzen, die nicht populationsbezogen definiert sind: Mit dem Standardmeßfehler und dem daraus konstruierten Konfidenzintervall für den wahren Wert kann man die Meßgenauigkeit des Tests charakterisieren, mit dem Standardschätzfehler die Vorhersagegenauigkeit. Wenn die Voraussetzung der Homoskedastizität (gleiche Meßgenauigkeit bzw. Vorhersagegenauigkeit in allen Skalenbereichen) erfüllt ist, so sind Standardmeßfehler und Standardschätzfehler von der Verteilung der wahren Werte unabhängig. Sie sind allerdings in ihrer numerischen Bedeutung an die verwendete Skala gebunden, die eine Rohpunktskala oder auch eine populationsspezifisch definierte Normskala (siehe 2.5) sein kann.

2.4 Die Rolle der Normalverteilung in der Testtheorie

Bei der Testkonstruktion strebt man meist eine Normalverteilung der Testrohwerte an. Hinweise, wie man die Itemschwierigkeiten zusammenstellen soll, um gute Aussichten auf normalverteilte Testrohwerte zu haben, findet man bei Lienert (1961). Die Normalverteilung hat vielfach inhaltliche Plausibilität: Bei Eigenschaften, die von sehr vielen Determinanten abhängen, ist es plausibel, daß Extremwerte selten, mittlere häufig zustande kommen. Zudem verteilen sich verschiedene körperliche Merkmale (z.B. die Körpergröße) annähernd normal. Vor allem aber hat die Normalverteilung besonders einfache statistische Eigenschaften: In multivariaten Normalverteilungen sind alle Regressionen linear und homoskedastisch, die Abweichungen von der Regressionslinie verteilen sich wieder normal. Konstruiert man Tests so, daß sie sich in der einschlägigen Bezugspopulation (z.B. einem Altersjahrgang) normal verteilen, so kann man erwarten, daß man zwischen diesen Tests einfache Beziehungen findet. Mathematisch gesehen folgt zwar aus der Normalverteilung jedes einzelnen Tests nicht die multivariate Normalverteilung als gemeinsame Verteilung, praktisch gesehen sind jedoch dafür günstige Voraussetzungen geschaffen. Sind Paralleltests bivariat normalverteilt, so sind die Meßfehler normalverteilt, und ihre Varianz ist in allen Skalenbereichen gleich groß. Damit sind die Voraussetzungen für die Konfidenzintervalle mit dem Standardmeßfehler (z.B. Konfidenzgrenzen für den wahren Wert, Angabe von kritischen Differenzen usw.) erfüllt. Analoges gilt bei bivariater Normalverteilung von Test und Kriterium: Zur Vorhersage kann die lineare

Beispiel 2.2: Populationsabhängigkeit von Reliabilitäts- und Validitätskoeffizienten: Angaben aus der Handanweisung zum Zahlen-Verbindungs-Test (ZVT) nach Oswald & Roth (1978).

Der ZVT ist ein nicht-verbaler Kurz-Intelligenz-Test, der ab einem Alter von 8 Jahren verwendet werden kann. Die Testhandanweisung enthält vielfältige Angaben zu Reliabilität und Validität. So werden auf S. 16 u.a. Testwiederholungs-Reliabilitäten nach Schularten getrennt und in der Gesamtstichprobe angegeben (zwei Stichproben zu je 96 Schülern, Alter 14 Jahre).

Testwiederholung	Sonderschüler	.86
nach 6 Wochen	Hauptschüler	.94
n = 96	Realschüler	.84
	Gymnasiasten	.94
	Insgesamt	.95
Testwiederholung	Sonderschüler	.95
nach 6 Monaten	Hauptschüler	.90
	Realschüler	.87
	Gymnasiasten	.85
	Insgesamt	.97

In beiden Fällen sieht man, daß die nach Schulart getrennt berechneten Koeffizienten niedriger liegen als der Koeffizient für die Gesamtstichprobe. Das ist aufgrund der Varianzeinschränkung in den Teilstichproben gegenüber der Gesamtstichprobe auch zu erwarten. (Angaben zu den Streuungen findet man auf S. 47: Die Streuungen des IQ liegen in den einzelnen Schularten zwischen 10 und 13, während sie in einer alle Schularten umfassenden repräsentativen Stichprobe 15 beträgt). Die Abhängigkeit der Korrelationskoeffizienten von der Homogenität oder Heterogenität der Stichprobe zeigt sich auch bei den Validitäts-Koeffizienten. Auf S. 20-21 der Handanweisung werden u.a. Korrelationen des ZVT zum Prüf-System für Schul- und Bildungsberatung nach Horn (PSB, 1969), dem Intelligenz-Struktur-Test nach Amthauer (IST, 1955, 1970) für verschiedene Stichproben (u.a. altersrepräsentative und nach Schularten getrennte Stichproben) mitgeteilt. Die Ergebnisse werden wie folgt zusammengefaßt (S. 21): “Die korrelativen Zusammenhänge in den repräsentativ gestalteten Stichproben eines Umfangs zwischen N = 45 und N = 126 zwischen PSB und ZVT sowie IST und ZVT variieren zwischen r = -.69 und r = -.80. (....). In den homogeneren und damit bezüglich einer hypothetischen Intelligenznormalverteilung varianzbeschnittenen Stichproben fielen die beobachteten Zusammenhange etwas geringer aus: Sie variieren bei Stichproben zwischen N = 24 und N = 100 zwischen r = -.40 und r = -.83 und liegen im Durchschnitt bei r = -.50” (Anmerkung: die negativen Vorzeichen ergeben sich daraus, daß beim ZVT die Bearbeitungszeit gemessen wird, also hohe Werte schlechten Leistungen entsprechen).

Regression verwendet werden, die Vorhersagegenauigkeit ist in allen Skalenbereichen gleich gut und kann mit dem Standardschätzfehler angegeben werden.

Auch wenn aus den genannten Gründen eine Normalverteilung der Testwerte als wünschenswert gilt, so ist es doch nicht sinnvoll, in jedem Fall Normalverteilung erreichen zu wollen. Wenn z.B. die Einstellungen zu einem Themenbereich polarisiert sind, so ist eine zweigipfelige Verteilung zu erwarten, und es macht wenig Sinn, einen Gipfel im Mittelbereich erzwingen zu wollen. Ähnliches gilt, wenn nach Symptomen von Verhaltensstörungen gefragt wird, die in der Normalpopulation selten sind. Hier wird die Verteilung schief sein (die meisten Probanden weisen keine oder nur einige wenige Symptome auf), ohne daß das ein Mangel des Tests wäre.

In anderen Fällen können schiefe Verteilungen dadurch bedingt sein, daß bei der Messung eines Merkmals, das sonst normalverteilt ist (z.B. Werte in Intelligenztests), nur Items hoher Schwierigkeit (rechtsschiefe Verteilung) oder nur Items niedriger Schwierigkeit (linksschiefe Verteilung) herangezogen wurden. In solchen Fällen sollte der Test um entsprechende (schwerere oder leichtere) Aufgaben ergänzt werden. Problematisch wäre es in einem solchen Fall, Normalverteilung nur durch eine nachträgliche Transformation der Rohwertskala künstlich herzustellen. Höchst wahrscheinlich würde die Meßgenauigkeit nach der Normalisierung ungleich, und zwar in den künstlich gedehnten Skalenbereichen schlechter sein.

Schiefe Verteilungen treten auch bei Schulnoten und anderen Einschätzungsskalen häufig auf. Bei der Berechnung von Korrelationen stellt sich dann die Frage, ob die Skalen zunächst transformiert werden sollen, um für alle Variablen eine Normalverteilung zu erhalten. Praktisch bedeutet das, daß den gleichen Abständen auf der Notenskala ungleiche Abstände auf der transformierten Skala zugeordnet werden. Ob dadurch tatsächlich Linearität der Regression erreicht wird und die in den Daten vorhandenen Zusammenhänge besser beschrieben werden, bleibt jedoch im Einzelfall zu prüfen.

2.5 Die Normierung von Testwerten

Testergebnisse liegen zunächst in Form von Rohwerten (Anzahl richtig gelöster Aufgaben, Anzahl der Ja-Antworten, für die Lösung benötigte Zeit usw.) vor. Um die Interpretation des Testwerts eines Probanden zu erleichtern, ist es in der Regel nützlich, ihn mit den Testwerten anderer Probanden (Probanden gleichen Alters, gleicher oder anderer Schulbildung, verschiedene Berufsgruppen usw.) zu vergleichen. Deshalb ist es wünschenswert, daß der Testautor für möglichst viele verschiedene Populationen Vergleichsdaten zur Verfügung stellt. Solche Vergleichsdaten werden in Form von Normentabellen angegeben.

Wohl am verbreitetsten ist die Normierung anhand von Altersjahrgängen, die zunächst im Zusammenhang mit Intelligenztests eingeführt wurde. Um Normentabellen für eine bestimmte Altersstufe zu erstellen, muß zunächst aus dieser Altersstufe eine repräsentative Stichprobe gezogen und die Verteilung der Testrohwerte festgestellt werden. Angenommen, die Rohwerte verteilen sich normal, so kann man sie in z-Werte umrechnen und ihnen die entsprechenden Prozentränge der Standard-Normalverteilung zuordnen. Ein z-Wert gibt an, wieviele Streuungseinheiten ein Proband über dem Durchschnitt liegt.

[2.17] $$z = \frac{X - \mu}{\sigma}$$

X = Testrohwert
p,cj = Mittelwert und Streuung der Testrohwerte in der Normierungspopulation, z.B. einem Altersjahrgang

Die z-Werte haben den Mittelwert Null und die Streuung Eins. Von da aus kann man zu einer Skala mit beliebig festgelegtem Mittelwert und beliebig festgelegter Streu-

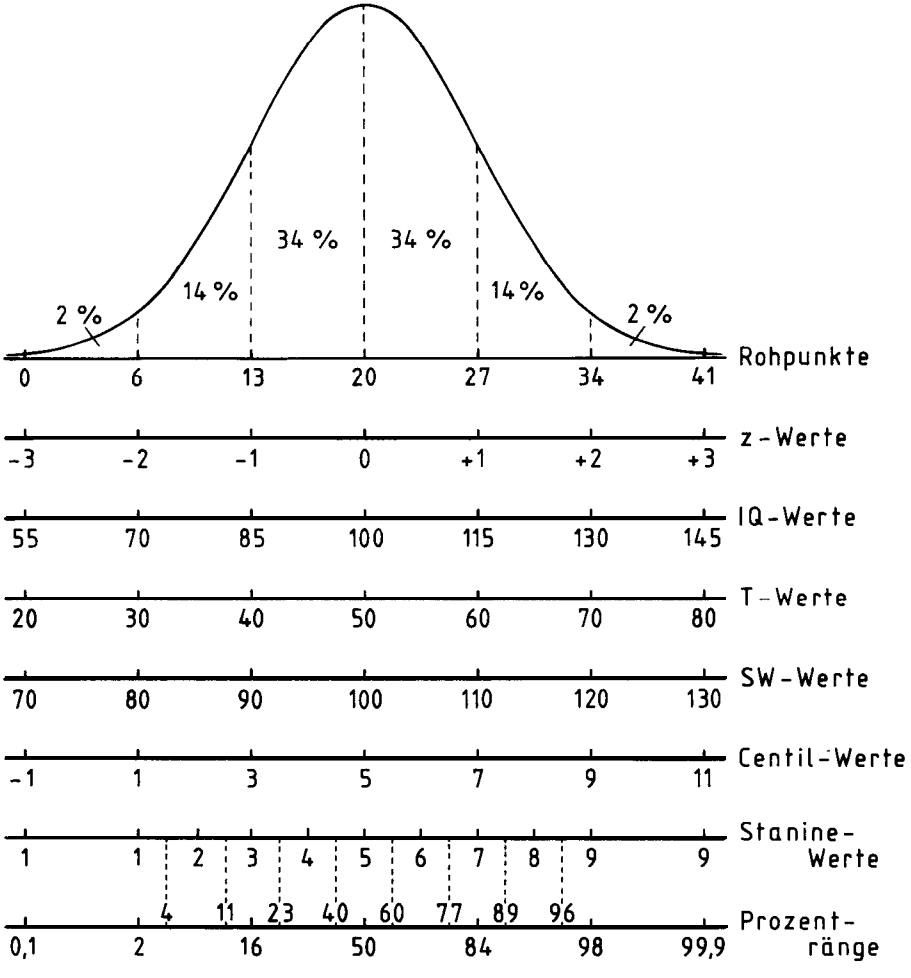


Abbildung 2.2: Normalverteilte Testrohwerte mit Mittelwert $\mu = 20$ und Streuung $\sigma = 7$ und verschiedene gebräuchliche Normskalen

ung übergehen, indem man den z-Wert mit der gewünschten Streuung multipliziert und den gewünschten Mittelwert addiert. Im folgenden sind einige gebräuchliche Arten der Normierung angegeben:

Intelligenz-Quotienten:	IQ= 100 + 15z
T-Werte:	T = 50 + 10z
Standardwerte, Z-Werte:	Z = SW = 100 + 10z
Centilwerte:	C = 5 + 2 z

Die Umrechnung von Rohwerten in z-Werte nach Formel [2.17] und der darauf folgende Übergang zu einer der angegebenen Normskalen sind lineare Transformationen (lineare Transformationen erlauben, daß eine Meßwertreihe mit einer Zahl a multipliziert wird und daß eine Zahl b hinzuaddiert wird). Lineare Transformationen lassen die Intervalleigenschaften einer Skala unverändert: Abstände, die auf der einen Skala gleich groß sind, sind auch auf der anderen Skala gleich groß. Abbildung 2.2 zeigt normalverteilte Rohwerte und verschiedene Normskalen (z, IQ, T, SW, C), die daraus durch lineare Transformationen hervorgehen.

In der untersten Reihe von Abbildung 2.2 werden den Rohwerten Prozentränge zugeordnet. Ein Prozentrang gibt zu jedem Rohwert X an, wieviel Prozent der Probanden in einer Population einen Rohwert kleiner/gleich X erzielen. Hat z.B. ein Proband einen Punktwert erreicht, dem ein Prozentrang von 80 entspricht, so heißt das, daß in der Population 80% der Probanden einen niedrigeren, höchstens gleichen, und 20% der Probanden einen höheren Punktwert erreichen. In Abbildung 2.2 sieht man, daß bei normalverteilten Rohwerten Prozentränge keine lineare Transformation der Rohwerte sind: Demselben Rohpunktunterschied entspricht im Mittelbereich ein grosser, an den Skalenenden ein kleiner Unterschied im Prozentrang. Die Prozentrangskala dehnt also bei einer Normalverteilung die Abstände im Mittelbereich und staucht sie an den Enden. Prozentränge haben aber den Vorteil, daß sie eine anschauliche, auch dem Laien leicht verständliche Bedeutung haben.

Die Stanine-Skala baut auf der Prozentrang-Skala auf. Sie hat insgesamt 9 Stufen (das Wort “Stanine” steht kurz für “Standard nine”). Sie ordnet den Prozenträngen wie folgt Skalenwerte zu:

Prozentrang	Stanine	Relative Häufigkeit
o - 4		4 %
über 4 - 11	2	7 %
über 11 - 23	3	12 %
über 23 - 40	4	17 %
über 40 - 60	5	20 %
über 60 - 77	6	17 %
über 77 - 89	7	12 %
über 89 - 96	8	7 %
über 96 - 100	9	4 %

Die Zusammenfassung der Prozentränge zu den Stufen der Stanine-Skala erfolgt so, daß die Stanine-Werte (von der Vergröberung auf nur 9 Skalenwerte abgesehen) normalverteilt sind, wobei der Mittelwert der Normalverteilung auf 5 und die Streuung auf 2 festgesetzt ist. Die Stanine-Skala entspricht somit einer Centil-Skala, bei der die Centilwerte unter 1 dem Wert 1 und die Centilwerte über 9 dem Wert 9 zugeschlagen werden.

Die Umrechnung des Testergebnisses von Rohwerten in Normwerte dient dazu, das Testergebnis eines Probanden relativ zu den Leistungen in einer Vergleichspopulation, der Normpopulation, anzugeben. Bei Intelligenztests ist eine Normierung bezogen auf Altersstufen üblich. Bei anderen Tests mögen andere Bezugspopulationen sinnvoller sein: bei Schulleistungstests z.B. die Schüler der entsprechenden Schulstufe und Schulart, usw. Die Erhebung der Testnormen ist die aufwendigste Phase in der Testkonstruktion. Hier kommt es darauf an, wirklich repräsentative Stichproben zu ziehen. Jede Verzerrung in der Normierungsstichprobe führt zu entsprechenden Fehlern bei der Beurteilung der einzelnen Probanden, die später mit dem Test untersucht werden: War die Normstichprobe verglichen mit der Population zu "gut", so wird hinterher der einzelne Proband zu "schlecht", weil zu streng beurteilt. Ist in der Normstichprobe die Varianz reduziert (weil z.B. im Streben nach Repräsentativität möglichst "durchschnittliche" Schulen in die Normstichprobe aufgenommen wurden), so erscheint später die Stellung des Probanden extremer (positiv oder negativ) als sie in der Population tatsächlich ist. Da die Datenerhebung für eine Testnormierung sehr aufwendig ist und deshalb viele Tests in diesem Punkt Mängel aufweisen, wird man bei der Beurteilung der Qualität eines Tests auf die Qualität der Normen besonders zu achten haben. Es gibt allerdings auch verschiedene Einsatzbereiche von Tests, wo keine Normen benötigt werden. Deshalb wird die Normierung auch nicht zu den Hauptgütekriterien gerechnet. Wenn es z.B. darum geht, die Probanden mit den höchsten oder niedrigsten Testleistungen zu selegieren, so genügen Testrohwerte. Dasselbe gilt für viele Fragestellungen in der Forschung, wenn z.B. Mittelwerte verschiedener Gruppen verglichen oder Korrelationen mit Testleistungen berechnet werden sollen. Wenn alle Probanden aus derselben Population stammen, so daß der Übergang von Rohwerten zu Normwerten lediglich eine lineare Transformation der Meßwerte darstellt, so hat eine solche Transformation keinerlei Einfluß auf die Höhe der Korrelationen oder die Signifikanz von Mittelwertsunterschieden, ist also überflüssig. Testnormen sind vor allem für die beratende Diagnostik von Interesse, indem sie helfen, das Testergebnis des einzelnen Probanden in Relation zu verschiedenen Vergleichspopulationen richtig einzuordnen.

Einführende Literatur:

Belser, H. (1967). *Testentwicklung. Verfahren und Probleme der Entwicklung von Gruppen-Intelligenztests, dargestellt am Beispiel der Frankfurter Analogietests.* Weinheim: Beltz.

Lienert, G.A. (1991). *Testaufbau und Testanalyse.* 5.Auflage. Weinheim: Psychologie Verlags Union.

Erläuterungen zu Begriffen aus der Statistik und Testtheorie findet man auch bei:

Kriz, J. & Lisch, R. *(1988). Methoden-Lexikon **für** Mediziner, Psychologen, Soziologen.* München: Psychologie Verlags Union.

Weiterführende Literatur:

Lord, F. M. & Novick, M. R. (1968). *Statistical theories **of** mental test scores.* Reading, Mass.: Addison-Wesley.

Fischer, *G.* H. (1974). *Einführung in die Theorie psychologischer Tests.* Bern: Huber.

3. Die Interpretation von Testbatterien

1. Wie bildet man aus mehreren Untertests einen Gesamtstandardwert? Was ist der Unterschied zwischen dem Gesamtstandardwert und der mittleren Profilhöhe?
2. Wie groß muß die Differenz zweier beobachteter Testwerte mindestens sein, damit mit hinreichender Sicherheit ausgeschlossen werden kann, daß sie nur durch Meßfehler zustande gekommen ist?
   Wie verteilen sich die Differenzen zwischen zwei Untertests in der Population?
   Wie häufig wird eine bestimmte Differenz überschritten?
   Wann ist eine Untertest-Leistung verglichen mit den übrigen erwartungswidrig?
3. Welche Probleme treten auf, wenn das Testprofil eines Probanden mit einem Gruppenprofil (z.B. Berufsgruppenprofil) als Anforderungsprofil verglichen werden soll?

Vorstrukturierende Lesehilfe

Viele Tests mit umfassendem Geltungsanspruch, wie z.B. allgemeine Intelligenztests, Schulleistungstests, aber auch umfassendere Fragebogen, bestehen aus einer Reihe von Untertests. Bei der Auswertung kann jeder Untertest für sich betrachtet und der Untertestwert mit den Normdaten in Beziehung gesetzt werden. Die standardisierten Untertestwerte können dann miteinander verglichen werden, um Stärken und Schwächen des Probanden zu beschreiben, oder auch um besondere Diskrepanzen (z.B. als Hinweis auf pathologische Ausfälle) zu diagnostizieren. Darüber hinaus wird gewöhnlich ein Gesamttestwert gebildet, der das Testergebnis insgesamt möglichst gut repräsentieren soll. Im folgenden soll zunächst der Gesamttestwert betrachtet werden (3.l), danach werden typische Fragestellungen behandelt, die auftreten, wenn im Sinne einer Profilinterpretation die Leistungen eines Probanden in verschiedenen Untertests verglichen werden (3.2). Im letzten Abschnitt (3.3) werden Probleme behandelt, die bei der Interpretation von Gruppenprofilen als Anforderungsprofile auftreten.

3.1 Zum Gesamttestwert

Um den Gesamttestwert zu bilden, werden zunächst die Ergebnisse aus den einzelnen Untertests addiert. Sofern die Rohpunkte in den einzelnen Untertests in etwa die gleiche Varianz haben, können einfach die Rohpunkte zu einer Rohpunktsumme addiert werden. Wenn jedoch die Rohpunktvarianzen stark unterschiedlich sind, weil

Beispiel 3.1: Berechnung des Mittelwerts aus den standardisierten Untertestwerten eines Probanden ("mittlere Profilhöhe") und seines Gesamtstandardwerts

Zwei Untertests seien beide auf SW-Einheiten (Mittelwert 100, Streuung 10) standardisiert. Ihre Korrelation sei p = 0.7. Das Gesamttestergebnis soll ebenfalls auf SW-Einheiten standardisiert werden (= Gesamtstandardwert GSW) Ein Proband habe im ersten Test einen Wert von $SW_1 = 110$, im zweiten Test einen Wert von $SW_2 = 120$. Man berechne seine mittlere Profilhöhe und seinen GSW.
Lösung:
Mittlere Profilhöhe: (110 + 120)/2 = 115
Gesamtstandardwert: Dazu benötigen wir zunächst die SW-Summe. Sie beträgt 110 + 120 = 230. Um diesem Wert einen z-Wert zuzuordnen, müssen wir Mittelwert und Varianz der SW-Summe berechnen:
Für den Mittelwert (Erwartungswert) erhält man:
$E(SW_1 + SW_2) = E(SW_1) + E(SW_2) = 100 + 100 = 200.$
Die Varianz der Summe erhält man als Summe der Varianzen plus Kovarianzen:

$$\sigma^2 (SW_1 + SW_2) = \sigma^2 (SW_1) + \sigma^2 (SW_2) + 2 \rho \cdot \sigma(SW_1) \sigma (SW_2)$$
$$= 100 + 100 + 2 \cdot 0.7 \cdot 10 \cdot 10 = 340$$

und für die Streuung:

$$\sigma(SW_1 + SW_2) = 18.4$$

Damit erhalten wir für den z-Wert:

$$z = \frac{230 - 200}{18.4} = 1.6$$

Aus dem z-Wert errechnet sich dann der Gesamtstandardwert als
GSW= 100 + 10z = 100 + 10. 1.6 = 116
Verglichen mit der mittleren Profilhöhe (115) fällt der GSW (116) extremer aus. Indem man denselben Rechengang mit anderen Werten von ρ durchführt, kann man sich leicht überzeugen, daß der Unterschied zwischen GSW und mittlerer Profilhöhe umso deutlicher wird, je niedriger die Untertestkorrelation ist.

z.B. Aufgabenzahl und Art der Aufgabenbewertung bei den einzelnen Untertests recht verschieden ist, würden bei einer Addition von Rohpunkten die Untertests mit größeren Varianzen das Gesamttestergebnis entsprechend stärker bestimmen. Formal betrachtet: Die Varianz der Rohpunktsumme ist die Summe der Varianzen und Kovarianzen. (Erwartungswert und Varianz von Summen und gewichteten Summen von Zufallsvariablen sind ausführlich bei Stange, 1970, Kapitel 5.4, kürzer bei Lord & Novick, 1968, Kapitel 4.6 behandelt). Ein Test mit großer Rohwertvarianz hat einen entsprechend großen Anteil an der Varianz der Rohpunktsumme. Diese ungleiche Gewichtung der Untertests wäre an entsprechend ungleichen Korrelationen der einzelnen Untertests zum Gesamttestwert abzulesen. Will man eine solche ungleiche Gewichtung vermeiden, so muß man bei ungleichen Varianzen die Untertestwerte zunächst standardisieren (z.B. auf SW-Einheiten), und erst diese standardisierten Werte werden zu einer Gesamtsumme (SW-Summe) aufaddiert.

Im nächsten Schritt müssen für den Summenwert (Rohpunktsumme oder Summe von standardisierten Werten, z.B. SW-Summe) Mittelwert und Streuung berechnet werden. Sind diese Werte bekannt, so kann man auf dem bereits dargestellten Weg (Berechnung von z-Werten, anschließende Transformation auf den gewünschten Mittelwert und die gewünschte Streuung; vgl. Kapitel 2.5) zu standardisierten Gesamttestwerten übergehen.

Bei oberflächlicher Betrachtung könnte man meinen, der standardisierte Gesamttestwert müßte dem Durchschnitt der standardisierten Untertestwerte entsprechen. Das ist jedoch nicht der Fall. Der Durchschnitt der standardisierten Untertestergebnisse heißt *"mittlere Profilhöhe"* und ist vom standardisierten Gesamttestwert zu unterscheiden. Die mittlere Profilhöhe hat zwar denselben Mittelwert wie der standardisierte Gesamttestwert, die Varianz hängt aber außer von den Untertestvarianzen auch von den Untertestkorrelationen ab. Außer in dem unrealistischen mathematischen Spezialfall von zu Eins korrelierenden Untertests ist die Varianz der mittleren Profilhöhe kleiner als die Varianz des standardisierten Gesamttestwerts. Das bedeutet praktisch, daß der standardisierte Gesamttestwert immer etwas extremer (d.i. bei überdurchschnittlichen Werten höher, bei unterdurchschnittlichen Werten niedriger) ausfällt als die mittlere Profilhöhe. Beispiel 3.1 illustriert diesen Unterschied am Spezialfall von nur zwei Subtests. Bei nur zwei Subtests würde man zwar kaum von einem "Profil" sprechen (der Intelligenz-Struktur-Test von Amthauer, 1970, z.B., der eine Profilauswertung vorsieht, besteht aus zehn Untertests), der begriffliche Unterschied zwischen mittlerer Profilhöhe und Gesamtstandardwert läßt sich aber auch mit zwei Subtests rechnerisch demonstrieren. Bei zehn Untertests ist der Rechengang analog, nur langwieriger.

3.2 Zur Interpretation von Untertest-Differenzen

a) Die Berechnung kritischer Differenzen

Wenn man die standardisierten Untertestwerte eines Probanden vorliegen hat, so liegt es nahe, sie untereinander zu vergleichen und die Differenzen im Sinn besonderer Stärken oder Schwächen zu interpretieren. Dabei stellt sich zunächst die Frage, ob eine Differenz groß genug ist, damit mit hinlänglicher Sicherheit ausgeschlossen werden kann, daß sie nur durch Meßfehler zustande gekommen ist. Diese Frage wird durch die Angabe der kritischen Differenz beantwortet: Zunächst berechnet man für jeden der beiden Tests die Fehlervarianz wie in Formel [2.7] angegeben, als

$$\sigma^2 (F) = \sigma^2 (X) \cdot (1 - Rel)$$

Wenn die Meßfehler der beiden Tests jeweils mit dem Erwartungswert Null und den Varianzen $\sigma^2 (F_1)$ und $\sigma^2 (F_2)$ unabhängig normalverteilt sind, so sind die Meßfehler der Differenzen ebenfalls mit Erwartungswert Null normalverteilt und ihre Varianz ist:

[3.1] $$\sigma^2 (F_{Diff}) = \sigma^2 (F_1) + \sigma^2 (F_2)$$

Die bei einem Probanden gefundene Differenz ist bei $\alpha = 0.05$ signifikant, wenn

sie außerhalb des 95%-Bereichs der Meßfehlerverteilung liegt. Das ist der Fall, wenn sie die kritische Differenz

[3.2] $$D_{krit} = 1.96\ \sigma\ (F_{Diff})$$

dem Betrag nach übersteigt.

Bisweilen werden zur Profilinterpretation auch komplexere Maße empfohlen. So z.B. empfiehlt Amthauer (1970) den Durchschnitt aus den Untertests "Analogien" und "Zahlenreihen" mit dem Durchschnitt aus den Untertests "Gemeinsamkeiten" und "Rechenaufgaben" zu vergleichen. In anderen Fällen mag es interessant sein, die Differenz von einem einzelnen Subtest zum Durchschnitt aller anderen Subtests zu betrachten. Auch bei solchen komplexer zusammengesetzten Differenzmaßen stellt sich die Frage nach der kritischen Differenz, die zu überschreiten ist, damit eine Erklärung durch Meßfehler mit hinreichender Sicherheit auszuschließen ist. Eine aus mehreren Untertests gebildete Differenz hat die allgemeine mathematische Form:

$$D = a_1 X_1 + a_2 X_2 + \ldots a_k X_k \qquad \text{mit } \Sigma a_i = 0$$

Dabei sind die Koeffizienten a_i so zu wählen, daß die gewünschte Differenz. ausgedruckt wird. Will man z.B. bei fünf Untertests die Differenz zwischen den ersten beiden zum Durchschnitt der anderen drei Untertests bilden, so sind die a_i wie folgt bestimmt:

$$D = (X_1 + X_2) \cdot \frac{1}{2} - (X_3 + X_4 + X_5) \cdot \frac{1}{3}$$

$$\text{d.h.} \quad a_1 = a_2 = \frac{1}{2} \qquad a_3 = a_4 = a_5 = -\frac{1}{3}$$

Aufgrund der Unabhängigkeit der Meßfehler gilt dann für die Fehlervarianz der Differenz:

[3.3] $$\sigma^2(F_D) = a_1^2\sigma^2(F_1) + a_2^2\sigma^2(F_2) + \ldots a_k^2\sigma^2(F_k)$$

im vorliegenden Beispiel also:

$$\sigma^2(F_D) = \frac{1}{4}\sigma^2(F_1) + \frac{1}{4}\sigma^2(F_2) + \frac{1}{9}\sigma^2(F_3) + \frac{1}{9}\sigma^2(F_4) + \frac{1}{9}\sigma^2(F_5)$$

Die kritische Differenz ergibt sich dann, wie bereits als Formel [3.2] angegeben, als

$$D_{krit} = 1.96\ \sigma(F_D) \qquad \text{bei } \alpha = 0.05$$

Wenn die beim Probanden gefundene Differenz dem Betrag nach größer ist als die kritische Differenz, so ist sie signifikant.

Beispiel 3.2 illustriert die Berechnung von kritischen Differenzen sowohl für den Fall von zwei einzelnen Tests als auch für den Vergleich von zwei Subtestgruppen.

b) Die Häufigkeitsverteilung von Differenzen

Wenn man bei einem Probanden eine signifikante Differenz gefunden hat, so kann man weiter fragen, wie häufig eine solche oder größere Differenz in der Population vorkommt. Diese Frage interessiert, wenn man wissen will, ob es sich vielleicht um eine ungewöhnlich große Differenz handelt, der besondere diagnostische Bedeutung zu-

Beispiel 3.2: Die Berechnung von Kritischen Differenzen

1) Kritische Differenz zwischen zwei Untertests: Der Intelligenz-Struktur-Test IST 70 von Amthauer (1970) enthält u.a. die Untertests RA (eingekleidete Rechenaufgaben) und ZR (Zahlenreihen Fortsetzen).
Ein Schüler hat bei RA einen SW=120 und bei ZR einen SW=105 erreicht. Ist der Unterschied groß genug, daß mit 95% Sicherheit ausgeschlossen werden kann, daß er nur durch Meßfehler zustande gekommen ist?

Lösung: Zunächst sind die Fehlervarianzen der beiden Tests nach Formel [2.7] zu bestimmen. Dazu benötigt man die Varianzen und Reliabilitäten der beiden Untertests. Die Varianz von Standardwerten (SW) ist definitionsgemäß 100. Die Reliabilitäten entnimmt man der Handanweisung: Bei Testwiederholung nach einem Jahr mit einer Parallelform wurde für RA eine Reliabilität von .86, für ZR eine Reliabilität von .75 gefunden. Daraus ergibt sich:
$\sigma^2 (F_{RA}) = 100(1 - .86) = 14$ und
$\sigma^2 (F_{ZR}) = 100(1 - .75) = 25$
Danach errechnet man die kritische Differenz nach Formel [3.2]:
$D_{krit} = 1.96 \sqrt{14 + 25} = 12.2$
Die bei unserem Probanden gefundene Differenz beträgt 120 - 105 = 15. Damit ist die kritische Differenz dem Betrag nach überschritten und die gefundene Differenz ist bei $\alpha = .05$ signifikant. Das heißt, eine so große oder noch größere Differenz kommt bei gleichen wahren Werten meßfehlerbedingt in weniger als 5% der Fälle zustande. Die Differenz kann interpretiert werden.

2) Berechnen der kritischen Differenz für den Vergleich von Subtestgruppen: Ein Schüler schneidet im Mathematikunterricht bei den geometrischen Aufgaben regelmäßig schlechter ab als bei den rechnerischen. Von daher wird erwartet, daß er im Intelligenztest bei Tests des räumlichen Vorstellens schlechter abschneidet als bei numerisch-mathematischen Aufgabenstellungen.
Der IST 70 enthält außer den o.g. zwei rechnerischen Tests RA und ZR mit den Reliabilitäten .86 und .75 auch zwei Tests zum räumlichen Vorstellen, nämlich FA = Figurenauswahl und WÜ = Würfelaufgaben mit den Reliabilitäten .69 und .65. Der Schüler hat folgende Standardwerte (SW) erzielt:
RA: 120, ZR: 105, FA: 90, Wü: 92
Ist der Unterschied zwischen den rechnerischen und den räumlichen Tests groß genug, damit mit 95%iger Sicherheit ausgeschlossen werden kann, daß er nur durch Meßfehler zustande gekommen ist ?

Lösung: Sein Durchschnitt aus den beiden rechnerischen Test ist demnach (120 + 105)/2 = 112.5, sein Durchschnitt aus den beiden räumlichen Tests (90 + 92)/2 = 91. Sein Durchschnitt für die rechnerischen Tests liegt also um 21.5 Einheiten höher als für die räumlichen Tests (gefundene Differenz).
Um die kritische Differenz zu berechnen, berechnet man zunächst die Fehlervarianzen für alle Tests nach Formel [2.7]:

	RA	ZR	FA	Wü
σ^2 (F)	14	25	31	35

Das Differenzmaß, für das die Fehlervarianz bestimmt werden soll, lautet:

D = (RA + ZR)/2 - (FA + Wü)/2 = .5 RA + .5 ZR -.5FA -.5Wü

Die Fehlervarianz für das Differenzmaß ergibt sich dann nach Formel [3.3] als

$\sigma^2 (F_D) = .5^2 \cdot 14 + .5^2 \cdot 25 + (-.5)^2 \cdot 31 + (-.5)^2 \cdot 35 = 26.25$
$\sigma(F_D) = 5.12$
Daraus erhält man die kritischen Differenz als für $\alpha = .05$
$D_{krit} = 1.96 \cdot 5.12 = 10.04$
Die beim Probanden gefundene Differenz von 21.5 ist dem Betrag nach größer als die kritische Differenz und damit bei $\alpha = .05$ signifikant.

Anmerkung: Man sieht, daß trotz der relativ niedrigen Reliabilitäten der beiden räumlichen Tests die kritische Differenz für den Vergleich der Mittelwerte aus den beiden Subtestgruppen sogar etwas niedriger ausfällt als die kritische Differenz zum Vergleich der beiden reliableren Einzeltests RA und ZR. Dieses Ergebnis ist typisch und kommt dadurch zustande, daß beim Mitteln der Testwerte auch die Meßfehler gemittelt werden, wodurch die Fehlervarianz des Mittelwerts reduziert wird.

kommt. Hat man es mit nur zwei gleich standardisierten bivariat normalverteilten Subtests zu tun, so läßt sich die Verteilung der Differenzen leicht errechnen. Die Differenzen sind wieder normalverteilt. Der Mittelwert der Verteilung ist Null, die Varianz ergibt sich als

[3.4] $$\sigma^2 (D) = \sigma^2 (X_1) + \sigma^2 (X_2) - 2\, \rho(X_1 X_2)\sigma(X_1)\, \sigma(X_2)$$

Kennt man somit Mittelwert und Varianz der Differenzen-Verteilung, so kann man zu jeder beliebigen Differenz D den zugehörigen z-Wert berechnen und dann mithilfe der Tabelle für die Standard-Normalverteilung feststellen, wie häufig dieser Wert überschritten wird. Bei Differenzmaßen, die aus mehr als zwei Subtests bestehen, ist analog vorzugehen. Die Berechnung der Varianz ist allerdings langwieriger, weil sie von den Kovarianzen aller beteiligten Untertests untereinander abhängt. Beispiel 3.3 illustriert den Rechenvorgang für den Fall von zwei Subtests.

Ist die Normalverteilungsvoraussetzung nicht gegeben, so bleiben obige Aussagen über Mittelwert und Varianz der Differenzenverteilung gültig: Bei gleich standardisierten ‘Fests ist der Mittelwert der Differenzen Null und die Varianz der Differenzen ergibt sich nach Formel [3.4]. Die genaue Häufigkeitsverteilung der Differenzen kann aber nicht über Normalverteilungstabellen bestimmt werden, sondern muß empirisch ermittelt werden (Berechnen der Differenz für jeden Probanden aus einer repräsentativen Stichprobe, Aufstellen der Häufigkeitsverteilung).

Wenn die Werte jedes einzelnen Subtests normal verteilt sind, so folgt daraus zwar mathematisch gesehen noch nicht zwingend die bivariate Normalverteilung und damit die Normalverteilung der Differenzen, in der Praxis wird man aber bei normal verteilten Subtests auch mit einer Normalverteilung der Differenzen rechnen können.

Beispiel 3.3: Häufigkeit von Differenzen in der Population

Die in Beispiel 3.1 genannten Tests RA und ZR korrelieren zu .58. Ein Proband hat im Test RA um 15 SW-Einheiten besser abgeschnitten als in ZR. Ist eine solche Differenz ungewöhnlich groß?
Lösung: Um diese Frage zu beantworten, berechnen wir die Häufigkeit, mit der eine SW-Differenz von 15 oder mehr SW-Einheiten in der Population vorkommt. Nimmt man an, daß RA und ZR für die entsprechende Altersstufe bivariat normalverteilt sind, so ist die Differenz ebenfalls normalverteilt. Der Mittelwert dieser Normalverteilung ist Null, die Varianz erhält man nach Formel [3.4] als

Einer Differenz von D = 15 entspricht demnach folgender z-Wert einer Standardnormalverteilung:

$$z = \frac{15 - 0}{9.16} = 1.64$$

Einer Tabelle für die Standardnormalverteilung entnimmt man, daß außerhalb des Bereichs ± 1.64 noch 11% der Fälle liegen. D.h.: In der Population besteht bei 11% der Probanden eine Differenz zwischen RA und ZR, die 15 SW-Einheiten oder noch mehr beträgt. Die bei dem Probanden gefundene Differenz ist also nicht extrem selten.

c) Abweichungen von Regressions-Schätzwerten

Statt zu fragen, ob eine Differenz häufig oder selten vorkommt, kann man bei Vorliegen entsprechender Hypothesen auch gezieltere Fragen stellen. Wenn z.B. bekannt ist, daß sich eine organische Hirnschädigung speziell auf einen Test X_2 auswirkt, nicht aber auf X_1, kann man gezielt fragen, ob der Proband in X_2 eine signifikant schlechtere Leistung erbringt, als aufgrund seiner Leistung in X_1 zu erwarten wäre. Dazu nimmt man eine Regressionsschätzung von X_2 aus X_1 vor und betrachtet die Abweichung von diesem Schätzwert. Bivariate Normalverteilung vorausgesetzt, ist die Regression von X_2 auf X_1 linear und die Abweichungen von der Regression verteilen sich wieder normal mit dem Mittelwert Null und der Varianz

[3.5] $$\sigma^2 (X_2 - X^*_2) = \sigma^2 (X_2) \cdot (1 - \rho^2 (X_1 X_2))$$

X^*_2 = geschätzter Wert für X_2 (vgl. Formel [2.10])

Hat man für den Probanden die Regressionschätzung von X_2 aus X_1 vorgenommen und die Abweichung $X_2 - X^*_2$ bestimmt, so kann man berechnen, wie häufig eine solche oder größere Abweichung in der Population der Gesunden vorkommt. Ist sie sehr selten, so wird man ihr entsprechende diagnostische Relevanz beimessen. Der Einfachheit wegen war bisher angenommen worden, daß X_2 aus nur einem anderen Test x_1 geschätzt wird. X_1 kann aber auch ein komplexeres Maß, z.B. ein Gesamttestwert sein. Das Argument läuft dann analog.

Beispiel 3.4 illustriert den Rechenvorgang bei der Regressionschätzung und der Berechnung der Abweichung des Probanden von dieser Regressionsschätzung.

Beispiel 3.4: Berechnung der Abweichung vom Regressions-Schätzwert

Ein Wortschatztest (X_1) und ein Test für kurzfristiges Behalten (X_2) sind bei Gesunden auf SW-Einheiten normiert und korrelieren zu 0.50. Bei einem Probanden wird vermutet, daß das kurzfristige Behalten krankheitsbedingt gestört ist. Der Proband hat in Test X_1 einen SW von 120, in X_2 einen SW von 90 erreicht.
Wie häufig kommt es in der Population der Gesunden vor, daß jemand, der in X_1 einen SW von 120 hat, in X_2 nur 90 oder weniger hat?
Lösung: Wenn die Tests X_1 und X_2 bei Gesunden bivariat normalverteilt sind, so kann der Durchschnitt von X_2 für Probanden mit $X_1 = 120$ mittels linearer Regression berechnet werden. Die Regressionsgleichung zur Schätzung von X_2 aus X_1 (vgl. Formel [2.10]) lautet:

$$X^*_2 = \beta X_1 + \alpha$$

mit $\beta = \rho(X_1 X_2)\, \sigma(X_2)/\sigma(X_1)$ und $\alpha = E(X_2) - \beta E(X_1)$

Für das vorliegende Beispiel ergibt sich also:

$$\beta = 0.5 \cdot 10/10 = 0.5 \text{ und } \alpha = 100 - 0.5\ 100 = 50$$

$$X^*_2 = 0.5\, X_1 + 50 = 0.5 \cdot 120 + 50 = 110$$

Probanden mit $X_1 = 120$ erreichen also in X_2 im Durchschnitt einen Wert von 110. Die Varianz um diesen Durchschnitt beträgt nach Formel [3.5]:

$$\sigma^2 (X_2 - X^*_2) = 100(1 - 0.5^2) = 75;\ \sigma (X_2 - X^*_2) = 8.7$$

Das heißt: Betrachtet man nur gesunde Probanden, die in $X_2 = 120$ haben, so verteilen sich deren Werte in X_2 normal um den Mittelwert 110 mit einer Streuung von 8.7. Unser Proband hat einen Wert von $X_2 = 90$. Dem entspricht folgender z-Wert:

$$z = (90 - 110)/\ 8.7 = -2.29$$

Einer Tabelle für die Standardnormalverteilung entnimmt man, daß dieser z-Wert nur in 1.1% der Fälle unterschritten wird. D.h.: Unter gesunden Probanden mit einem Standardwert von 120 in Test X_1 haben nur 1.1% in Test X_2 einen Standardwert von 90 oder weniger. Der vorliegende Befund ist also bei Gesunden sehr selten, was die Vermutung einer krankheitsbedingten Störung stützt. Wenn entsprechende Angaben auch für die Population der Kranken zur Verfügung stehen, kann die Rechnung analog auch für diese Population durchgeführt werden.

Zur Wahl eines Entscheidungskriteriums bei der Interpretation von Differenzen

Da man in der psychologischen Statistik gewohnt ist, das Signifikanzniveau auf $\alpha = .05$ festzulegen, liegt es nahe, eine entsprechende Forderung auch für die individuelle Diagnostik zu erheben, also die Nullhypothese gleicher wahrer Werte erst zu verwerfen, wenn die kritische Differenz für das 5%-Niveau überschritten ist. Zieht man zusätzlich in Betracht, daß bei einer Vielzahl von Untertests jeder mit jedem verglichen wird, also eine Vielzahl von Signifikanztests durchgeführt wird, so liegt es nahe, noch strengere Anforderungen zu stellen, um das Gesamtrisiko, einen oder mehrere Alpha-Fehler zu machen, nicht über 5% ansteigen zu lassen. Eine solche

Strategie würde allerdings dazu führen, daß nur sehr große Differenzen interpretiert werden und bei einem Großteil der Probanden die Nullhypothese beibehalten wird. Damit würde bei der großen Zahl von Personen, bei denen sich die wahren Werte der Subtests unterscheiden (bei denen also die Alternativhypothese zutrifft), dieser Unterschied nicht diagnostiziert. Da auch das Übersehen von Unterschieden für die diagnostische Praxis eine Fehlentscheidung bedeutet, muß eine Entscheidungsstrategie gewählt werden, die beide Fehlerrisiken angemessen berücksichtigt. Im folgenden sollen anhand eines Zahlenbeispiels die Fehlerraten für verschiedene Entscheidungs-Strategien berechnet werden: Bei Strategie A soll jede beobachtete Differenz interpretiert werden, bei Strategie B nur Differenzen, die größer sind als eine halbe Streuungseinheit, bei Strategie C nur Differenzen, die größer sind als die bei $\alpha = .05$ berechnete kritische Differenz.

Wir nehmen an, die Tests X_1 und X_2 seien bivariat normalverteilt, die Meßfehlerverteilungen normal und homoskedastisch. Beide Tests sollen auf SW-Einheiten (Mittelwert = 100, Streuung = 10) standardisiert sein, und beide sollen eine Reliabilität von 0.9 haben. Die Korrelation der beiden Tests betrage $\sim(X_1\ X_2) = .50$.

Die Verteilung der beobachteten Differenzen $X_1 - X_2$ und der wahren Differenzen $T_1 - T_2$ ist dann ebenfalls bivariat normal. Wir berechnen zunächst die Korrelation der beiden Variablen $X_1 - X_2$ und $T_1 - T_2$, um dann mithilfe von Tabellen für die bivariate Normalverteilung die Fehlerraten zu ermitteln. Als Fehler bewerten wir alle Fälle, bei denen wir die beobachtete Differenz gegenüber dem Probanden interpretiert haben, also aus $X_1 > X_2$ auf $T_1 > T_2$ geschlossen haben, wohingegen $T_1 \leq T_2$ richtig ist, d.h. in Wahrheit kein Unterschied besteht oder der Unterschied in die entgegengesetzte Richtung geht (analog: aus $X_1 > X_2$ auf $T_1 > T_2$ geschlossen, während $T_1 \leq T_2$ richtig ist).

Um die Korrelation zwischen den beobachteten und den wahren Differenzen zu bestimmen, benutzen wir einen Satz aus der Testtheorie (er ergibt sich als Spezialfall aus Formel 2.12a, wenn man dort für Y den Testwert X einsetzt), wonach die Korrelation zwischen beobachteten und wahren Werten gleich der Wurzel aus der Reliabilität ist. Bezogen auf einen einzelnen Test X lautet der Satz:

$$\rho(X\ T_x) = \sqrt{Rel(X)},$$

angewendet auf Differenzen ergibt sich:

$$\rho(X_1 - X_2\ ,\ T_1 - T_2) = \sqrt{Rel(X_1 - X_2)}$$

Um die Reliabilität der Differenz zu bestimmen (Anteil der wahren Varianz an der beobachteten Varianz), berechnen wir zunächst die beobachtete Varianz, dann die Fehlervarianz und schließlich die wahre Varianz als Differenz zwischen beobachteter und wahrer Varianz:

$$\sigma^2(X_1 - X_2) = \sigma^2(X_1) + \sigma^2(X_2) - 2\,\rho_{12}\,\sigma(X_1)\sigma(X_2)$$

$$= 100 + 100 - 2 \cdot 0.5 \cdot 10 \cdot 10 = 100$$

$$\sigma^2(F_1) = \sigma^2(X_1)(1 - Rel) = 100(1 - .9) = 10; \quad \text{analog: } \sigma^2(F_2) = 10$$

$$\text{und } \sigma^2(F_1 - F_2) = \sigma^2(F_1) + \sigma^2(F_2) = 10 + 10 = 20$$

$$\sigma^2 (T_1 - T_2) = 100 - 20 = 80$$

$$Rel(X_1 - X_2) = 80/100 = .80$$

Damit erhalten wir für die Korrelation zwischen beobachteten und wahren Differenzen:

$$\rho(X_1 - X_2, T_1 - T_2) = \sqrt{Rel(X_1 - X_2)} = \sqrt{.80} = .9$$

Nachdem wir die Korrelation berechnet haben, können wir nun die von Taylor & Russell (1939) publizierten Tabellen benutzen, um für die drei in Betracht gezogenen Selektionsstrategien die Fehlerraten zu bestimmen.

Taylor & Russell (1939) stellten im Zusammenhang mit der Frage nach der Nützlichkeit des Testeinsatzes in der betrieblichen Personalselektion erstmals Überlegungen zu den Fehlerraten verschiedener Selektionsstrategien an: Sie nahmen an, daß Testwerte und Berufserfolg bivariat normalverteilt sind. Ab einem bestimmten kritischen Kriteriumswert auf der Skala des Berufserfolgs gilt der Bewerber als "erfolgreich", darunter als "nicht erfolgreich". Der Anteil der Erfolgreichen in der Grundgesamtheit der Bewerber wird als *Grundquote* bezeichnet. Würde man per Zufall auswählen, so würde sich ein der Grundquote entsprechender Anteil als erfolgreich erweisen. Die Selektion wird nun aber mit Hilfe des Tests durchgeführt: Es werden alle Bewerber, die einen bestimmten kritischen Testwert überschritten haben, aufgenommen. Der Anteil der Aufgenommenen an den Bewerbern ist die *Selektionsquote*. Sind Grundquote, Selektionsquote und Test-Kriteriums-Korrelation bekannt, so kann man aus den Taylor-Russell-Tafeln die *Trefferquote* entnehmen. Darunter versteht man den Anteil der Erfolgreichen unter den Aufgenommenen. Die Nützlichkeit des Testeinsatzes wird dann danach beurteilt, wie weit die Trefferquote (Selektion mit Hilfe des Tests) über der Grundquote (Selektion nach Zufall) liegt. Weiterführende Überlegungen zur Nutzenmaximierung findet man bei Cronbach & Gleser (1965), Kurzdarstellungen bei Wottawa & Hossiep (1987) und bei Noack & Petermann (1988).

Zu Strategie A (jede beobachtete Differenz wird interpretiert): Um die Taylor-Russell-Tafeln zu benutzen, braucht man die Korrelation zwischen dem Kriterium (in unserem Anwendungsfall die wahre Differenz) und dem Test (allgemeiner gesagt: der korrelierenden Variablen, nach der die Selektion durchgeführt wird; in unserem Fall ist das die beobachtete Differenz). Die Korrelation beträgt also in unserem Fall .9. Weiter braucht man die Grundquote: In unserem Fall ist der "kritische Kriteriumswert" $T_1 - T_2 = 0$. Er wird von 50% der Probanden überschritten. Die Grundquote ist demnach 50%. Weiter braucht man die Selektionsquote: Der Anteil der Probanden mit $X_1 - X_2 > 0$ ist 50% (für diese Probanden machen wir die "Vorhersage" $T_1 - T_2 > 0$). Die Selektionsquote ist demnach 50%. Mit diesen Angaben kann man nun den Taylor-Russell-Tafeln die Trefferquote entnehmen: Sie beträgt 86%. Das heißt: 86% der "ausgewählten" Probanden (der Probanden mit $X_1 > X_2$) überschreiten den kritischen Kriteriumswert (haben wahre Werte $T_1 > T_2$), 14% erreichen den kritischen Kriteriumswert nicht (für sie gilt: $T_1 < T_2$). Da sich die Überlegung völlig analog für eine Selektion nach $X_1 - X_2 > 0$ und den kritischen Kriteriumswert $T_1 - T_2 = 0$ anstellen läßt, kann man zusammenfassend feststellen, daß Strategie A (jede Differenz wird interpretiert) bei einer Reliabilität der Differenz von .8 zu einer Fehlerrate von 14% führt.

Zu Strategie B : Nur Differenzen, die größer sind als eine halbe Streuungseinheit (= 5 SW-Einheiten), werden interpretiert: Korrelation und Grundquote sind gleich wie bei Strategie A. Die Selektionsquote ist niedriger, da nur über Probanden mit $X_1 - X_2 > 5$ eine Aussage gemacht wird. Der Anteil der Probanden mit $X_1 - X_2 > 5$ beträgt 31% (zum Rechengang siehe Beispiel 3.3). Den Taylor-Russell-Tafeln entnimmt man nun bei einer Selektionsquote von 31% eine Trefferquote von 97% und eine Fehlerrate von 3%.

Betrachtet man nun alle Probanden mit $X_1 > X_2$ (sie machen 50% der Grundgesamtheit aus), so verteilen sie sich wie folgt:

Keine Diagnose erstellt, weil $X_1 - X_2 < 5$ SW-Einheiten 19%
Diagnose "$T_1 > T_2$" erstellt 31%
davon richtige Diagnosen 30%
davon falsche Diagnosen 1%

Eine analoge Rechnung läßt sich für die Probanden mit $X_1 > X_2$ erstellen. Faßt man beide Gruppen zusammen, so erhält man folgendes Bild:

Keine Diagnose erstellt, weil $|X_1 - X_2| < 5$ SW-Einheiten 38%
Richtige Diagnosen 60%
Falsche Diagnosen 2%
Anteil der richtigen Diagnosen an den erstellten 97%

Zu Strategie C: Nur Differenzen, die die bei $\alpha = .05$ errechnete kritische Differenz überschreiten, werden interpretiert: Die kritische Differenz bei $\alpha = .05$ beträgt $D_{krit} = 8.8$ (zum Rechengang siehe Beispiel 3.1). Der Anteil der Probanden mit $X_1 - X_2 > 8.8$ beträgt 19% (Rechengang siehe Beispiel 3.3). Den Taylor-Russell-Tafeln entnimmt man bei einer Selektionsquote von nunmehr 19% eine Trefferquote von 99%. Die Probanden mit $X_1 > X_2$ verteilen sich damit wie folgt auf die Entscheidungen:

Keine Diagnose erstellt, weil $X_1 - X_2 < 8.8$ SW-Einheiten 31%
Diagnose "$T_1 > T_2$ " erstellt 19%
davon richtige Diagnosen 18.8%
davon falsche Diagnosen 0.2%

Die Rechnung für Probanden mit $X_1 > X_2$ ist wieder analog durchzuführen. Zusammen gefaßt über alle Probanden ergibt sich folgende Verteilung der Entscheidungen:

Keine Diagnose erstellt, weil $|X_1 - X_2| < 8.8$ 62%
Richtige Diagnosen 37.6%
Falsche Diagnosen. 0.4%
Anteil der richtigen Diagnosen an den erstellten 99%

Vergleicht man nun die drei Diagnose-Strategien A, B und C, so sieht man, wie mit zunehmender Enthaltung bei der Diagnosestellung der Anteil der richtigen Diagnosen an den erstellten steigt. Aber selbst Strategie A, bei der jede Differenz interpretiert wird, hat keine extrem hohen Fehlerraten. Wenn man vermutet, daß die diagnostische Praxis in etwa Strategie B folgt, so erscheint das durchaus rational: Der Anteil der erstellten Diagnosen ist relativ hoch, die Fehlerrate bei den Diagnosen noch akzeptabel.

Die genauen Zahlenwerte hängen natürlich von den in diesem Beispiel getroffenen Annahmen über die Reliabilitäten der beiden Tests und die Korrelation zwischen

den Tests ab. Die hier gemachten Annahmen sind aber nicht unrealistisch. Bei höheren Reliabilitäten und niedrigerer Korrelation zwischen den Tests fallen die Ergebnisse noch günstiger aus.

3.3 Zur Interpretation von Gruppenprofilen als Anforderungsprofile

Bei umfassenden Testbatterien, die verschiedene Intelligenz- oder Leistungsbereiche getrennt erfassen (z.B. IST 70 von Amthauer, 1970; LPS von Horn, 1983), werden als Interpretationshilfe u.a. die Durchschnittsprofile bestimmter Berufsgruppen, Studienrichtungen usw. angeboten. Für den Praktiker liegt es nun nahe, diese Durchschnittsprofile als Anforderungsprofile zu interpretieren: Wer Arzt werden will, sollte in allen Untertests die Durchschnittswerte der Ärzte erreichen oder überbieten.

Gegen so einfache Schlußfolgerungen sind jedoch Bedenken verschiedener Art anzumelden: Mittelwerte allein sagen noch nichts über die Relevanz des Untertests für den Berufserfolg aus. Das wird leicht erkennbar, wenn man ein offensichtlich irrelevantes Merkmal betrachtet: Würde man bei Ärzten die Körpergröße als "Untertest" miterheben, so würde man vermutlich zu dem Ergebnis kommen, daß sie ungefähr dem Durchschnitt der Bevölkerung entsprechen, also auf einer SW-Skala einen Mittelwert von etwa 100 haben. Das bedeutet aber nicht, daß körperlich Kleine nicht gute Ärzte werden könnten. Der Durchschnittswert ist hier kein Anforderungswert. Um die Relevanz eines Untertests für den Berufserfolg zu belegen, sind weitere Angaben nötig: z.B. Korrelationen des Tests mit Kriterien des Berufserfolgs, Angaben über Unterschiede zwischen erfolgreichen und erfolglosen Teilnehmern an einer Ausbildung, usw. Solche Daten sind freilich schwer zu erheben. Wo sie fehlen, könnte schon die Angabe der Varianz zusätzlich zum Mittelwert hilfreich sein: Wenn die Varianz bei erfolgreichen Vertretern des Berufs groß ist, ist zu vermuten, daß diesen Merkmal für den Berufserfolg nicht allzu kritisch ist.

Auch wenn man die Relevanz eines Untertests für einen bestimmten Beruf als gegeben unterstellt, bleibt die Frage offen, ab welchem Testwert man einen Probanden als geeignet betrachten soll. Wenn man fordert, daß der Durchschnitt der erfolgreichen Berufsvertreter erreicht werden muß, so bedeutet das, daß man nach dem Test rund die Hälfte der faktisch erfolgreichen Berufsvertreter ausscheiden würde (bei symmetrischen Verteilungen liegen genau 50% über dem arithmetischen Mittel, bei schiefen Verteilungen sind es je nach Art der Schiefe mehr oder weniger als 50%). Ein solches Kriterium erscheint als zu hoch angesetzt. Bei der Interpretation der Eignungs-Untersuchungs-Batterie (EUB nach Engelbrecht, 1975; 1978) z.B., die an den Arbeitsämtern zur Berufsberatung verwendet wird, wird ein Proband als geeignet betrachtet, wenn er in den relevanten Untertests einen berufsbezogenen Stanine-Wert von mindestens 2 hat, also nicht zu den untersten 4% der Berufsgruppe gehört. Eine empirisch begründete Festlegung würde voraussetzen, daß man Testwerte erfolgreicher und nicht erfolgreicher Berufsvertreter zur Verfügung hätte.

Der Möglichkeit, aus Durchschnittsprofilen von Berufsvertretern Anforderungsprofile für Berufsanwärter zu gewinnen, dürften - abgesehen von den Schwierigkeiten der Datenerhebung - auch grundsätzlich Grenzen gesetzt sein: Sowohl die beruflichen Anforderungen selbst als auch die Selektionsbedingungen, die den Zugang zu

den Berufen regeln, sind zeitlichen Veränderungen unterworfen. Bei attraktiven Berufen und knappem Angebot an Ausbildungsplätzen wird eine starke Selektion stattfinden, bei einem Überhang an Plätzen und Nachwuchsmangel wird die Selektion entsprechend gering sein. Die Auswahlkriterien der Ausbildungsstätten (Betriebe, Schulen usw.) werden sich nur zum Teil mit den berufliche Anforderungen decken. Empirisch gefundene Unterschiede zwischen Berufsgruppen spiegeln diese Selektionsvorgänge wider, in denen sich berufliche Anforderungen, Auswirkungen der Arbeitsmarktlage, Auswirkungen von richtigen und irrigen Meinungen über Berufe usw. mischen. Allzu subtile Vergleiche zwischen dem Profil eines einzelnen Probanden und Berufsgruppenprofilen als Anforderungsprofilen erscheinen von daher als nicht angebracht. Die Gruppenprofile können jedoch als grobe Orientierungsmarken betrachtet werden und als solche durchaus hilfreich sein.

Kristof (1958) bietet Formeln an, mit denen man prüfen kann, ob überhaupt ein Profil vorliegt. Bezogen auf einen einzelnen Probanden lautet die Nullhypothese: Die wahren Werte des Probanden sind in allen k Untertests gleich, die Unterschiede zwischen den beobachteten Testwerten sind nur durch Meßfehler zustande gekommen. Kritisch läßt sich einwenden, daß eine solche Nullhypothese extrem unplausibel ist und von daher eine statistische Überprüfung überflüssig erscheint.

Analoge Formeln werden von Kristof (1958) auch für die zufallskritische Prüfung von Gruppenprofilen vorgeschlagen. Hier ist zweierlei anzumerken: Zum einen ist auch hier die Gleichheit der durchschnittlichen wahren Werte in allen k Untertests eine sehr unplausible Hypothese. Zum anderen ist anzumerken, daß für die meisten praktischen Fragestellungen neben dem Meßfehler bei der Messung der einzelnen Probanden auch der Stichprobenfehler bei der Ziehung der Probanden aus der Grundgesamtheit von Belang ist. Wird nur die zufallskritische Absicherung gegen Meßfehler vorgenommen, so wird für die *spezielle vorliegende Stichprobe* von n Personen die Nullhypothese der Gleichheit der durchschnittlichen wahren Werte geprüft. Wenn nun aber z.B. das Berufsprofil der Bäcker zur Diskussion steht, so sollte die Nullhypothese $\mu_1 = \mu_2 = \ldots \mu_k$ geprüft werden, also die Gleichheit der Mittelwerte in der *Population* der Bäcker. Dazu sind entsprechende multivariate Verfahren (z.B. Hotellings T^2, siehe Anderson 1958, Kapitel 5.3.5) einzusetzen.

Ähnliches gilt, wenn zwei Gruppenprofile, z.B. 50 Bäcker versus 50 Maurer verglichen werden. Auch hier interessiert nicht die Frage, ob sich die durchschnittlichen wahren Werte dieser speziellen 50 Bäcker von denen jener speziellen 50 Maurer unterscheiden, sondern ob die Mittelwerte der beiden Grundgesamtheiten verschieden sind. Nur die erste Frage ist mit den Kristof-Formeln zu bearbeiten, die zweite, die Stichprobenfehler mit einbezieht, ist mit gängigen statistischen Verfahren für den multivariaten Mittelwertsvergleich (z.B. MANOVA) zu behandeln.

Eine weitere geläufige Fragestellung bezieht sich auf den Vergleich eines Einzelprofils mit einem Gruppenprofil. Auch hier erscheint die Nullhypothese, daß die wahren Werte des Einzelprofils den wahren Werten des Gruppenprofils genau gleichen (Differenz Null auf allen Skalen), so unwahrscheinlich, daß es keinen vemünftigen Grund gibt, sie überhaupt aufzustellen. Sinnvoller ist die Frage, wie der Abstand (oder umgekehrt ausgedrückt: die Ähnlichkeit) des Einzelprofils zum Gruppenprofil quantifiziert werden soll. Diese Frage stellt sich z.B., wenn man in der Berufsberatung die Ähnlichkeit des Profils eines Probanden zu verschiedenen Berufsgruppenprofilen ausdrücken will. Solche Fragen lassen sich mithilfe von Diskriminanzanalysen (siehe Kapitel 4.1) behandeln.

Zusammenfassung

(1) Um einen Gesamtstandardwert zu erhalten, wird zunächst eine Punktsumme (Rohpunktsumme oder Summe standardisierter Untertestwerte) gebildet. Diese wird wiederum standardisiert. Dieser Gesamtstandardwert ist nicht mit der mittleren Profilhöhe (= Durchschnitt aus den Untertest-Standardwerten) identisch, da letztere eine kleinere Varianz hat.

(2) Die Frage, ob zwei Untertestleistungen eines Probanden signifikant voneinander verschieden sind (Absicherung gegen eine Erklärung aus Meßfehlern), ist mithilfe der "kritischen Differenz" zu beantworten. Davon zu unterscheiden ist die Frage, wie häufig oder selten eine Differenz in einer Population überschritten wird, ob sie also in diesem Sinn "auffällig" ist. Diese Frage ist durch die Berechnung der Häufigkeitsverteilung der Differenzen zu beantworten. Eine weitere, davon zu unterscheidende Frage ist, ob ein bestimmter Untertest im Vergleich zu den übrigen Untertests erwartungswidrig niedrig (oder auch erwartungswidrig hoch) ausfällt. Diese Frage kann mithilfe einer Regressionsschätzung bearbeitet werden.

Bei der Frage, ab wann man den Unterschied zwischen zwei Untertestleistungen interpretieren soll, sind zwei Fehlerarten zu berücksichtigen: (1) Interpretieren von Unterschieden, die nur durch Meßfehler zustande gekommen sind, während sich die wahren Werte nicht unterscheiden oder der Unterschied sogar in die entgegengesetzte Richtung geht, und (2) Nicht-Diagnostizieren von vorhandenen Unterschieden. Die beiden Fehlerraten können für verschiedene Entscheidungsstrategien berechnet werden.

(3) Interpretiert man Durchschnittprofile, z.B. die Profile verschiedener Berufsgruppen, als Anforderungsprofile, so läuft man Gefahr, historische Zufälligkeiten des gesellschaftlichen Ist-Zustandes mit beruflichen Anforderungen zu vermengen. Über die Erhebung von Gruppenprofilen hinausgehend, sollte die berufsspezifische Relevanz der einzelnen Untertests und die Festlegung kritischer Anforderungsmarken empirisch begründet werden.

Einführende Literatur:

Lienert, G.A. (1991). *Testaufhau und Testanalyse* (5.Aufl.). Weinheim: Psychologie Verlags-Union.

Weiterführende Literatur:

Huber, H.P. (1973). *Psychometrische Einzelfalldiagnostik.* Weinheim: Beltz.

Abel, J. (1989). Profilanalysen in der Schulforschung. *Zeitschrift für Pädagogische Psychologie, 3,* 27 - 34.

4. Multivariate Verfahren im Dienst der Testtheorie

4.1 Verfahren zur Optimierung der Kriteriumsvorhersage: Multiple Regression und Diskriminanzanalyse

1. Wie kann man mehrere Tests zu einem Gesamtwert zusammenfassen, um eine möglichst genaue Kriteriumsvorhersage zu bekommen?
2. Wie genau fällt diese Vorhersage aus? Welche Tests können am ehesten weggelassen werden?

Vorstrukturierende Lesehilfe

Die meisten für die Praxis relevanten Kriterien, z.B. Schulerfolg oder Ausbildungserfolg, hängen von einer Vielzahl unterschiedlicher Bedingungen ab, wie Fähigkeiten, Kenntnissen, aber auch Interessen, Einstellungen und Erwartungen. Das Unterfangen, solch ein Kriterium mit einem einzelnen Test vorherzusagen, läßt von vornherein nur begrenzten Erfolg erwarten. Versucht man aber, mit unterschiedlichen Prädiktoren möglichst die gesamte Breite der Bedingungen zu erfassen, so stellt sich die Frage, wie diese unterschiedlichen Informationen relativ zueinander zu gewichten sind. Die Frage wird beantwortet, indem als Gesamtwert eine gewichtete Summe der Prädiktoren gebildet wird. Die Gewichtung wird bei einem quantitativ erfaßten Merkmal durch die multiple Regression, bei einem nicht-quantitativ erfaßbaren Kriterium (Zuordnung zu qualitativ verschiedenen Kategorien) durch die Diskriminanzanalyse bestimmt.

Als Prädiktoren können Informationen unterschiedlicher Art (Tests, Beurteilungen, Schulnoten, Alter u.a.) herangezogen werden. Wenn im folgenden von Tests als Prädiktoren die Rede ist, so ist das als Beispiel, nicht als Einschränkung zu verstehen.

4.1.1 Multiple Regression zur Maximierung der Kriteriumskorrelation

Die multiple Regression kann verwendet werden, um bei bereits feststehender Testauswahl die optimale Gewichtung zu finden, kann aber auch bei der Testauswahl selbst eingesetzt werden.

Wenn bereits feststeht, welche Tests X_1, X_2 . . X_p (z.B. die zehn Untertests des Intelligenz-Struktur-Tests von Amthauer, 1970) zur Vorhersage eines bestimmten Kriteriums Y (z.B. der Schulnote) verwendet werden sollen, so bestimmt die multiple Regressionsrechnung die Gewichte so, daß sich zwischen der gewichteten Summe der Tests und dem Kriterium eine maximale Korrelation ergibt. Man benötigt dazu die

Korrelationen aller Tests untereinander und mit dem Kriterium. Sofern nicht durch eine Standardisierung an der vorliegenden Stichprobe alle Variablen auf gleiche Mittelwerte und gleiche Streuung gebracht werden, benötigt man außerdem die Mittelwerte und Varianzen aller Variablen. Daraus lassen sich die optimalen Gewichte berechnen (das Verfahren kann hier nicht dargestellt werden. Es ist in jedem Lehrbuch über multivariate Statistik beschrieben; Literaturhinweise am Ende dieses Kapitels). Sie heißen *multiple Regressionsgewichte* (Beta-Gewichte). Hat man die Gewichte bestimmt, so wird der Kriteriumswert wie folgt geschätzt:

[4.1] $$Y^* = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_p X_p + \alpha$$

Y^* = geschätzter Kriteriumswert
$ß_1, ß_2, \ldots ß_p$ = multiple Regressionsgewichte (Beta-Gewichte)
α = Regressionskonstante

Die Regressionsgewichte hängen von den Kovarianzen der einzelnen Tests mit dem Kriterium, aber auch von den Kovarianzen der Tests untereinander ab. Jedes Hinzufügen weiterer Tests verändert in der Regel alle Regressionsgewichte. Die Regressionskonstante wird so bestimmt, daß Y und Y* den gleichen Mittelwert haben. Die Korrelation zwischen den Schätzwerten Y* und den tatsächlichen Kriteriumswerten Y heißt *multiple Korrelation* (R).

Berechnet man die multiplen Regressionsgewichte und die multiple Korrelation an derselben Stichprobe, so kommt es - insbesonders bei kleinen Stichproben und vielen Variablen - zu einer systematischen Überschätzung der multiplen Korrelation. Das liegt daran, daß die geschätzten Regressionsgewichte an die spezielle Stichprobe angepaßt werden, also z.B. bei 10 Tests immerhin 10 Parameter. Wenn die Stichprobe nur aus n = 10 Personen besteht, kann man mit 10 im nachhinein angepaßten Parametern in jedem Fall eine perfekte "Vorhersage" der Kriteriumswerte erzielen, selbst dann, wenn in der Grundgesamtheit keinerlei Zusammenhang bestehen sollte. Bei einer Stichprobe von n = 20 Personen und 10 Tests sind zwar mehr vorherzusagende Kriteriumswerte als anzupassende Parameter vorhanden, aber es wird immer noch zu einer deutlichen Überschätzung der multiplen Korrelation kommen. Dieser systematische Schätzfehler stellt allerdings kein grundsätzliches Problem dar, sondern kann mit Hilfe geeigneter Korrekturformeln behoben werden (Näheres dazu siehe Stevens, 1986, Kapitel 3.13).

Wenn noch nicht feststeht, welche Tests endgültig zur Kriteriumsvorhersage verwendet werden sollen, kann die Auswahl der Tests mit Hilfe einer schrittweisen multiplen Regression erfolgen. Bei der sogenannten "Vorwärts-Strategie" wird zunächst der Test gesucht, der die höchste Korrelation zum Kriterium hat. Unter den übrigen wird dann derjenige herausgesucht, der zusammen mit dem ersten die höchste multiple Korrelation (von nunmehr zwei Tests zum Kriterium) ergibt. Dieser Test kommt als zweiter in die Auswahl. Unter den verbleibenden wird wieder derjenige herausgesucht, der als dritter, zusammen mit den bereits ausgewählten Tests die höchste multiple Korrelation ergibt, usw. Das Verfahren wird abgebrochen, wenn der Zuwachs an multipler Korrelation, der sich bei Hinzunahme weiterer Tests erzielen läßt, nicht mehr als lohnend erscheint. Bei der sogenannten "Rückwärts-Strategie" berechnet man zunächst die multiple Korrelation unter Verwendung aller Tests und läßt dann schrittweise immer denjenigen Test weg, dessen Streichung zum geringsten Verlust an multipler Korrelation führt. Beide Strategien haben sich praktisch bewährt, bieten

aber mathematisch gesehen keine Garantie für eine optimale Lösung: Wählt man auf die beschriebene Art fünf von zehn Tests aus, so braucht Vorwärts- und Rückwärtsstrategie nicht zum selben Ergebnis zu führen und keine von beiden kann garantieren, daß es nicht eine noch bessere Fünfer-Kombination gibt. Rechenprogramme, wie z.B. SPSSX, bieten sowohl Vorwärts- als auch Rückwärtsstrategie, als auch gemischte Strategien an.

Wenn anhand von Stichprobendaten eine Untertest-Selektion stattgefunden hat, führt die Berechnung der multiplen Korrelation an denselben Daten der Erwartung nach zu einer systematischen Überschätzung der Güte der Vorhersage in der Population. Diese Überschätzung als Folge der Selektion geht über das hinaus, was durch die Anpassung der Regressionsgewichte bei feststehender Untertestauswahl bedingt ist. Die zu erwartende Überschätzung ist umso stärker, je kleiner die Stichprobe ist und je stärker selegiert wird. Dieses Problem ist nun nicht mehr durch Korrekturformeln lösbar, sondern erfordert eine sogenannte *"Kreuzvalidierung"*. Dafür müssen zwei unabhängige Datensätze zur Verfügung stehen (z.B. durch Zufallsaufteilung der Gesamtdaten in zwei Hälften). An dem einen Datensatz führt man die Untertestauswahl durch und bestimmt die multiplen Regressionsgewichte, an dem zweiten Datensatz wird die so gewonnene Schätzgleichung angewendet und die Korrelation zwischen Schätzwerten und Kriterium bestimmt. Diese kreuzvalidierte multiple Korrelation gibt dann eine unverzerrte Schätzung für die Güte der Vorhersage, die bei Anwendung der Schätzgleichung auf weitere Probanden erreicht werden wird. Beispiel (4.1) illustriert dieses Vorgehen.

Die in [4.1] angegebene Schätzformel geht davon aus, daß das Kriterium aufgrund einer gewichteten Summe der Testwerte vorhergesagt werden soll. Grundsätzlich ist es auch möglich, den in Formel [4.1] angegebenen Ansatz zu erweitern, indem man nichtlineare Ausdrücke (z.B. das Produkt zweier Testwerte, quadratische Funktionen der Testwerte usw.) hinzufügt, was allerdings in der Praxis kaum angewendet wird. Wenn Testwerte und Kriteriumswerte multivariat normalverteilt sind, ist die Regression des Kriteriums Y auf die Tests X_1, X_2 .. X_p linear, und der gemäß Formel [4.1] berechnete Schätzwert Y* liefert die bestmögliche Kriteriumsvorhersage, die aus den Tests zu erstellen ist.

Mit Hilfe der multiplen Regression scheint das Problem der Kriteriumsvorhersage optimal gelöst zu sein. Wenn es mit psychologischen Testbatterien gelänge, in einem Anwendungsbereich die wesentlichen Grunddimensionen individueller Unterschiede zu erfassen (z.B. für den Bereich der Schulleistungen die wesentlichen Intelligenzfaktoren, Interessens- und Einstellungsdimensionen), so könnte man die unterschiedlichen Kriterien (z.B. Noten in den einzelnen Schulfächern, Erfolg in verschiedenen Ausbildungsgängen) aus einer einheitlichen Testbatterie (allgemeiner: einem festen Satz von Prädiktoren) unter Verwendung der jeweils optimalen Gewichtung vorhersagen. Demgegenüber erscheint es zunächst überraschend, daß die multiple Regression in der Praxis so wenig genutzt wird: Kaum ein Testmanual enthält Berichte über Multiple-Regressions-Studien oder empfiehlt die Anwendung bestimmter Regressionsgewichte; lediglich einige Test-Kurzformen, die auf multiplen Regressions-Schätzungen des Gesamttestwerts beruhen, erfreuen sich größerer Verbreitung (z.B. WIP nach Dahl, 1972; WIPKI nach Baumett, 1973). Dafür dürften folgende Gründe verantwortlich sein:

Beispiel 4.1: Verwendung der multiplen Regression zur Vorhersage der Gesamttestleistung aus einer Kurzform

Der Hamburg-Wechsler-Intelligenztest für Kinder (HAWIK) besteht aus 10 Untertests. Baumett (1973) setzte sich zum Ziel, daraus eine Kurzform zu entwikkeln, die mit dem aus dem Gesamttest errechneten IQ möglichst hoch korrelieren soll. Daraus ergeben sich die Fragen, (a) welche der 10 Untertests verwendet werden sollen und (b) wie diese Untertests gewichtet werden sollen. Faßt man das Gesamttestergebnis als Kriterium Y auf und die Untertests als die Prädiktoren X_I bis X_{10}, so läßt sich diese Fragestellung mittels multipler Regression bearbeiten. Die folgende Darstellung des Vorgehens von Baumert (1973) ist vereinfacht und bezieht nur einen Teil der dort durchgeführten Analysen mit ein.

Als Daten standen die Testprotokolle von 614 Kindern zur Verfügung, die den ganzen Test bearbeitet hatten. In Hinblick auf die geplante Kreuzvalidierung wurde zunächst die Gesamtstichprobe nach dem Zufall in zwei Teilstichproben zu je 307 Testprotokollen aufgeteilt. Es wurde an jeder der beiden Teilstichproben getrennt eine schrittweise multiple Regression nach der Vorwärtsstrategie durchgeführt. Dabei ergab sich in beiden Stichproben nach Auswahl von vier Untertests eine hohe multiple Korrelation (.94 und .95).

Die in die Auswahl aufgenommenen Tests waren aber nicht genau dieselben: Nur drei der vier Tests (AW=Allgemeines Wissen, GF=Gemeinsamkeiten finden, BO=Bilder ordnen) waren in beiden Fällen in der Auswahl, als vierter Test tauchte einmal FL(=Figurenlegen), einmal MT(=Mosaiktest) auf. Aufgrund weiterer Gesichtspunkte, u.a. aufgrund der höheren Reliabilität von MT im Vergleich zu FL, wurde dann die Kombination AW,GF,BO,MT als Kurzform festgelegt.

Danach wurde die Schätzgleichung aufgestellt und kreuzvalidiert. Die Regressionsgewichte wurden zunächst an der einen Datenhälfte bestimmt, und dann an der anderen Datenhälfte angewendet, um die kreuzvalidierte Korrelation zu berechnen. Dabei zeigt sich nur eine minimale Schrumpfung der kreuzvalidierten gegenüber der an derselben Teilstichprobe berechneten multiplen Korrelation. Diese geringe Schrumpfung ist dem großen Stichprobenumfang von 2 mal 307 Personen zu verdanken.

Nach dieser Absicherung wurde als beste Schätzung der in der Population gültigen Regressionsgleichung die Regressionsgleichung aus den Gesamtdaten (n=614) berechnet. Sie lautet:

$$IQ^* = 33 + 1.84\,AW + 1.35\,GF + 1.41\,BO + 1.66\,MT$$

Für die Leistungen in den einzelnen Untertests sind dabei die jeweils erzielten Punkte (sog.“Wertpunkte”, die aus den Antworten des Probanden gemäß Testhandanweisung altersspezifisch zu bestimmen sind) einzusetzen. Die angegebene Formel schätzt dann aus den vier Untertests den IQ, den der Proband bei Vorgabe des ganzen Tests erhalten hätte.

Als Anmerkung kann man feststellen, daß sich die Regressionsgewichte für die vier Untertests nicht sehr stark unterscheiden. Das legt die Vermutung nahe, daß eine einfache ungewichtete Addition mit anschließender Transformation auf IQ-Einheiten keine wesentlich schlechteren Ergebnisse gebracht hätte.

(1.) Ein Hinzufügen oder Wegnehmen von Tests verändert in der Regel alle Regressionsgewichte. Eine multiple Regressionsschätzung ist also nur möglich, wenn genau die angegebene Testbatterie verwendet wird.

(2.) Die multiplen Regressionsgewichte ändern sich von Population zu Population. Alles, was auf die Korrelationen der Tests untereinander und die Korrelationen der Tests mit dem Kriterium Einfluß hat (insbesondere Selektionseinflüsse aller Art), beeinflußt auch die Regressionsgewichte. Eine multiple Regressionsgleichung ist also nur dann anzuwenden, wenn die zu beratenden Probanden aus derselben Population stammen, für die die Regressionsgewichte bestimmt wurden. Das aber erscheint vielfach als fraglich, zumal wenn die Regressionsstudie zeitlich und örtlich unter recht speziellen Bedingungen durchgeführt wurde.

(3.) Wenn die Tests gleich standardisiert sind und untereinander und mit dem Kriterium positiv korrelieren, liegt die multiple Korrelation nur wenig über dem Wert, den man bei einer einfachen gleichgewichtenden Addition erreicht (Wainer, 1976). Die einfache Addition hat aber Vorteile, wenn man daran denkt, daß das Testergebnis dem Ratsuchenden vermittelt werden muß: Das Abschneiden in den einzelnen Untertests sowie ein aus den Untertests gleich gewichtend errechneter Gesamtwert ist dem Probanden leicht verständlich zu machen. Die Gewichte der multiplen Regression können für den Probanden unplausibel sein und zu einer Ablehnung des darauf gegründeten Rates führen. Diese Gründe, zusammen mit dem erheblichen Datenaufwand, der mit dem Erstellen einer multiplen Regressionsgleichung verbunden ist, dürften wohl dafür verantwortlich sein, daß die multiple Regression in der Praxis nicht stärker zum Einsatz kommt.

4.1.2 Diskriminanzanalyse zur optimalen Trennung von Kriteriumsgruppen

Wenn das Kriterium nicht quantitativ erfaßt ist (wie z.B. Ausbildungserfolg, gemessen an den Abschlußnoten), sondern zwischen qualitativ verschiedenen Gruppen unterschieden werden soll (z.B. zwischen erfolgreichen Vertretern unterschiedlicher Berufsgruppen: zwischen mehreren klinischen Gruppen, o.ä.), kann eine Diskriminanzanalyse eingesetzt werden. Aus der Testbatterie wird dann - ähnlich wie bei der multiplen Regression - eine gewichtete Summe gebildet, wobei die Gewichte so gewählt werden, daß sich die Gruppen im Summenwert möglichst gut unterscheiden: Die Mittelwertsunterschiede zwischen den Gruppen sollen möglichst groß, die Varianz innerhalb der Gruppen möglichst klein sein. Die entsprechenden Gewichtszahlen heißen *Diskriminanzgewichte,* die mit den Diskriminanzgewichten aus den Testwerten gebildete gewichtete Summenvariable heißt *Diskriminanzfunktion.* Die Werte der einzelnen Probanden auf der Diskriminanzfunktion heißen *Diskriminanzwerte.* Die Diskriminanzgewichte hängen von den Mittelwerten der Gruppen in den Tests ab, aber auch von den Varianzen und den Kovarianzen der Tests untereinander, sowie von den relativen Anteilen, mit denen Vertreter der einzelnen Gruppen in der Stichprobe repräsentiert sind (bei einer Diskriminanzanalyse zur Unterscheidung zwischen Berufsgruppen vom Anteil der einzelnen Berufe an der Gesamtstichprobe).

Bei mehr als zwei Gruppen können mehrere Diskriminanzfunktionen gebildet werden, bei k Gruppen maximal k-l. Die erste wird so gewählt, daß sie eine bestmögli-

che Trennung der Gruppen (gemessen als Varianz zwischen den Gruppenmittelwerten relativ zur Varianz innerhalb der Gruppen) ermöglicht. Die Gewichte für die zweite Diskriminanzfunktion werden so gewählt, daß der resultierende Summenwert (zweite Diskriminanzfunktion) mit dem ersten unkorreliert ist. Unter dieser Restriktion wird wieder nach einer Gewichtung gesucht, die die Gruppen bestmöglich trennt. Die dritte Diskriminanzfunktion muß mit jeder der ersten beiden unkorreliert sein, usw. (zur rechnerischen Durchführung sowie zur Erweiterung des Ansatzes auf nichtlineare Funktionen sei auf die am Ende des Kapitels angeführten Lehrbücher verwiesen).

Kennt man ‘die Testwerte eines Probanden, so können daraus seine Werte in den Diskriminanzfunktionen berechnet werden. Wenn bestimmte Voraussetzungen erfüllt sind (die Testwerte sind in jeder Kriteriumsgruppe multivariat normalverteilt; die Kovarianzmatrizen sind gleich; die Grundraten, d. h. die Anteile, die die einzelnen Kriteriumsgruppen an der Gesamtpopulation ausmachen, sind bekannt), kann man daraus die bedingten Wahrscheinlichkeiten für die Zugehörigkeit zu den einzelnen Kriteriumsgruppen berechnen. Unter schwächeren Voraussetzungen kann man globale Ähnlichkeitsmaße verwenden, die die Nähe des Probanden zu den einzelnen Kriteriumsgruppen ausdrücken. Darauf aufbauend können verschiedene diagnostische Entscheidungsstrategien gewählt werden, nach denen die Probanden den Kriteriumsklassen zugeordnet werden: Man kann die Entscheidungsregel so wählen, daß einfach die Gesamtzahl richtig Klassifizierter maximiert wird, oder man kann verschiedene Arten von Fehlklassifikationen unterschiedlich stark gewichten und ein daraus abgeleitetes Nützlichkeitsmaß maximieren (Näheres dazu findet man bei Kallus & Janke, 1988).

Ähnlich wie bei der multiplen Regression kann man auch bei der Diskriminanzanalyse versuchen, durch schrittweises Hinzufügen von Tests eine möglichst sparsame Testbatterie zusammenzustellen, die eine möglichst gute Trennung der Kriteriumsgruppen erlaubt. Statt schrittweise hinzuzufügen (Vorwärtsselektion), kann man auch von einer gegebenen Testbatterie ausgehend schrittweise jeweils denjenigen Test weglassen, der am wenigsten zur Unterscheidung der Gruppen beiträgt (Rückwärtsselektion).

Bezüglich der Verallgemeinerbarkeit der Ergebnisse aus einer Diskriminanzanalyse sind dieselben Einschränkungen zu machen, wie bei einer multiplen Regression:

(1.) Ein Hinzufügen oder Wegnehmen von Tests verändert in der Regel alle Diskriminanzgewichte.

(2.) Ein Hinzunehmen oder Wegnehmen von Gruppen oder Verschiebungen in den relativen Anteilen der Gruppen an der Gesamtpopulation verändert in der Regel die Diskriminanzgewichte.

(3.) Wenn die Berechnung der Diskriminanzfunktionen und die Bestimmung der Vorhersagegenauigkeit (Prozent richtig klassifizierter Probanden) an derselben Stichprobe erfolgen, kommt es zu einer Überschätzung der Güte der Vorhersage. Das gilt in verstärktem Maß, wenn anhand derselben Daten eine Variablenselektion (s. oben) stattgefunden hat. Die Überprüfung der Vorhersagegenauigkeit sollte deshalb an einem neuen, unabhängigen Datenmaterial erfolgen (Kreuzvalidierung).

Eines der größten Forschungsprojekte im Bereich der angewandten Diagnostik, bei dem die Diskriminanzanalyse eingesetzt wird, dürfte die Entwicklung der maschinellen Auswertung der Testbogen in der Berufsberatung sein (Engelbrecht 1975; 1978). Bei der Bundesanstalt für Arbeit liegen aufgrund langjähriger Datensammlung inzwi-

schen für eine Vielzahl von Berufen Testwerte ehemaliger Ratsuchender vor, die inzwischen ihren Beruf erfolgreich ausüben. Aufgrund einer diskriminanzanalytischen Auswertung, die mit EDV.-Einsatz realisiert wird, ist es möglich, für jeden neuen Ratsuchenden die globale Ähnlichkeit zu den Vertretern der einzelnen Berufsgruppen als bedingte Wahrscheinlichkeit der Berufsgruppenzugehörigkeit anzugeben. Im Beratungsgespräch stellt sich allerdings das Problem, wie das Testergebnis an den Ratsuchenden zu vermitteln ist, so daß das Zustandekommen einer Empfehlung für den Probanden nachvollziehbar ist. Dazu sind Werte auf Diskriminanzfunktionen wenig geeignet. Deshalb wird zusätzlich für jede Berufsgruppe angegeben, in welchen Einzeltests (Leistungstests, Interessentests) der Proband relativ zu dieser Berufsgruppe sehr hohe oder sehr niedrige Werte aufweist, also vom durchschnittlichen Vertreter dieser Berufsgruppe stark abweicht. Sowohl Abweichungen nach oben (hohe Fähigkeiten oder Interessen in Bereichen, die für den Beruf nicht typisch sind) als auch nach unten (geringe Ausprägung von berufstypischen Interessen und Fähigkeiten) können auf Probleme hinweisen und Gegenstand des weiteren Beratungsgesprächs sein.

Zusammenfassung

Die Frage, wie mehrere Prädiktoren zu gewichten sind, um ein Kriterium bestmöglich vorherzusagen, wird bei quantitativ erfaßbaren Kriterien durch die multiple Regression, bei qualitativ et-faßbaren Kriterien durch die Diskriminanzanalyse beantwortet. Bei der multiplen Regression wird aus der gewichteten Summe der Prädiktoren ein geschätzter Kriteriumswert berechnet; Maß für die Güte der Vorhersage ist die multiple Korrelation.

Bei der Diskriminanzanalyse werden aus den Prädiktoren zunächst Werte auf Diskriminanzfunktionen berechnet; aus diesen wieder können bedingte Wahrscheinlichkeiten für die Zugehörigkeit zu den einzelnen Kriteriumsgruppen (oder andere Maße, die die Nähe des Probanden zu den einzelnen Kriteriumsgruppen ausdrucken) berechnet werden. Maß für die Güte der Vorhersage ist der Anteil richtig klassifizierter Probanden oder ein darauf aufbauendes Nützlichkeitsmaß. Die Ergebnisse sowohl einer multiplen Regression als auch einer Diskriminanzanalyse sind für die Personenpopulation und die spezielle Prädiktorenauswahl spezifisch und können in der Regel nicht darüber hinaus verallgemeinert werden.

Einführende Literatur:

Lehrbücher über multivariate statistische Verfahren:

Fahrmeir, L. & Harnerle, A. (Hrsg.) (1984). *Multivariate statistische Verfuhren.* Berlin: De Gruyter.

Hartung, J. & Elpelt, B. (1984). *Multivariate Statistik.* München: Oldenbourg.

Stevens, J. (1986). *Applied multivariate statistics for the social sciences.* Hillsdale: Lawrence Erlbaum.

Weiterführende Literatur:

Zur multiplen Regression:

Schubö, W., Haagen, K. & Oberhofer, W.(1983). Regressions- und kanonische Analyse. In J. Bredenkamp & H. Feger (Hrsg.), *Strukturierung und Reduzierung von Daten* (S.207-292). Göttingen: Hogrefe.

Zur Diskriminanzanalyse:

Krauth, J. (1983). Diskriminanzanalyse. In J.Bredenkamp & H.Feger (Hrsg.), *Strukturierung und Reduzierung von Daten* (S.293-350). Göttingen: Hogrefe.

Zu Klassifikations- und Entscheidungsstrategien in der Diagnostik:

Cronbach,L.J. & Gleser,G.C.(1965). *Psychological tests and personne1 decisions.* Urbana: University of Illinois Press.

Kallus, K.W. & Janke, W. (1988). Klassenzuordnung. In R.S. Jäger (Hrsg.), *Psychologische Diagnostik* (S. 131- 145). München: Psychologie Verlags Union.

Noack, H. und Petermann, F. (1988). Entscheidungstheorie. In R.S. Jäger (Hrsg.), *Psychologische Diagnostik* (S.241-253). München: Psychologie Verlags Union.

4.2 Faktorenanalyse zur Untersuchung der Konstruktvalidität

1) Von welchen Grundannahmen geht die klassische Faktorenanalyse aus?
2) Warum kommt es zwischen faktorenanalytischen Theorien zu keiner Entscheidung? Warum werden die Ergebnisse von Faktorenanalysen nicht mehr als funktional erklärende Theorien betrachtet?
3) Was unterscheidet konfirmatorische Faktoranalysen von exploratorischen? Welche Anwendungsmöglichkeiten bieten sich in der Testtheorie?

Vorstrukturierende Lesehilfe

Die Faktorenanalyse ist ein multivariates Verfahren, das mit der Geschichte der Theorienbildung im Bereich der Intelligenz- und Persönlichkeitsforschung und mit der Entwicklung psychologischer Tests besonders eng verbunden ist. Die Ideen der grossen Faktorentheoretiker wie Spearman, Guilford, Thurstone haben die Konzepte unserer heutigen Tests in Inhalt und Aufbau nachhaltig beeinflußt. Die heute im Gebrauch befindlichen psychometrischen Tests gehen überwiegend auf diese Vorbilder zurück.

Der ursprünglich hohe Anspruch, mit der Faktorenanalyse eine funktionale Analyse leisten zu können, z. B. die Grundfähigkeiten zu entdecken und zu messen, aus denen die menschlichen Intelligenzleistungen erklärbar würden, wird heute allerdings nicht mehr erhoben. Vielmehr betrachtet man heute die Faktorenanalyse als eine Methode, die geeignet ist, Korrelationsmuster überschaubarer zu machen und Interpretationsmöglichkeiten aufzuzeigen. Auch mit diesem reduzierten Anspruch kann sie zur Beantwortung der Frage nach der Validität eines Tests wertvolle Beiträge leisten. Im folgenden wird zunächst der Grundansatz der Faktorenanalyse dargestellt, dann werden die Hauptkritikpunkte wiedergegeben, die zu der erwähnten Rücknahme des Anspruchs geführt haben. Schließlich soll noch die konfirmatorische Faktorenanalyse als Weiterentwicklung der klassischen Faktorenanalyse bezüglich ihrer Anwendung auf testpsychologische Fragestellungen diskutiert werden.

4.2.1 Grundannahmen der Faktorenanalyse

Die folgende Darstellung orientiert sich am Modell mehrerer gemeinsamer Faktoren als dem allgemeinsten Ansatz. Andere Modelle lassen sich als Spezialfalle auffassen, die aus diesem Ansatz durch Zusatzannahmen hervorgehen (z. B. das Ein-Faktor-Modell durch die Annahme, es gebe nur einen gemeinsamen Faktor; das Hauptkomponenten-Modell durch die Annahme, die gesamte Testvarianz gehe auf gemeinsame Faktoren zurück). Eine umfassende Darstellung, die auch historische Aspekte miteinbezieht, gibt Pawlik (1971).

4.2.1.1 Die Grundgleichungen

Als Beispiel wollen wir annehmen, wir hätten die Korrelationen zwischen einer Vielzahl von Leistungstests (Intelligenztests, Schulleistungstests usw.) vorliegen. Es liegt

nahe, anzunehmen, daß diese Korrelationen dadurch zustande kommen, daß die Tests sich in ihren Anforderungen überschneiden, d.h. z.T. dieselben Fähigkeiten beanspruchen. Ziel der Faktorenanalyse ist es nun, solche mehreren Tests gemeinsamen Fähigkeiten zu definieren und ihr relatives Gewicht für die einzelnen Tests zu bestimmen. Allgemeiner gesprochen, besteht das Ziel der Faktorenanalyse darin, Korrelationen zwischen Variablen (hier: Leistungstests) auf gemeinsame Faktoren (= Dimensionen individueller Unterschiede, hier: Fähigkeiten) zurückzuführen und damit eine sparsame Interpretation der Korrelationen anzubieten.

Gemäß den Annahmen der Faktorenanalyse sind also für jede Testleistung mehrere Fähigkeiten (= Faktoren) erforderlich (z.B. zum Lösen eingekleideter Rechenaufgaben: Textverständnis, schlußfolgerndes Denken, Rechenfertigkeit), die sich mit unterschiedlichen Gewichten auf die Testleistung auswirken. Fähigkeiten (Faktoren), die von mehreren Tests (einer in einer Faktorenanalyse gemeinsam analysierten Testgruppe) beansprucht werden, heißen gemeinsame Faktoren, solche die nur in einem einzigen Test vorkommen, spezifische Faktoren. Darüber hinaus enthält jeder Test Meßfehler.

Gleichung [4.2] gibt an, wie die Testleistung einer Person in einem Test gemäß den Grundannahmen der Faktorenanalyse zustande kommt:

[4.2] $$z_{iv} = a_{i1} f_{1v} + a_{i2} f_{2v} + ...+ u_{iv}$$

Z_{iv} = Testwert der Person v im Test i, ausgedruckt in z-Werten, d.h. standardisiert auf Mittelwert 0 und Varianz 1.

a_{i1} = Gewicht, mit dem Faktor 1 die Testleistung im Test i bestimmt = Faktorladung des Tests i in Faktor 1.

a_{i2} = Faktorladung des Tests i in Faktor 2. Weitere Faktorladungen sind analog definiert.

f_{1v} = Faktorwert der Person v im ersten Faktor (individuelle Fähigkeitsausprägung in Faktor 1).

f_{2v} = Faktorwert der Person v im zweiten Faktor.

Weitere Faktorwerte für Person v sind analog definiert.

Die Faktorwerte sind für jeden Faktor auf den Mittelwert 0 und die Varianz 1 standardisiert.

U_{iv} = Durch die gemeinsamen Faktoren nicht erklärter Restanteil (englisch: uniqueness). Er enthält Einflüsse spezifischer Faktoren und Meßfehler und wird als von den gemeinsamen Faktoren unabhängig vorausgesetzt.

Aus dieser Grundgleichung ergibt sich, wie die Korrelation zwischen zwei Tests i und j zustande kommt: Die Leistung der Person v im Test j kann analog zerlegt werden (Gleichung [4.2a]):

[4.2a] $$z_{jv} = a_{j1} f_{1v} + a_{j2} f_{2v} + ...+ u_{jv}.$$

Betrachtet man nun die Kovarianz (= Korrelation, weil die Tests z-standardisiert sind) zwischen Test i und j, so sieht man, daß sie einerseits von den Gewichten (= Faktorladungen) abhängt, die die gemeinsamen Fähigkeiten für die beiden Tests haben, andererseits von den Korrelationen der Fähigkeiten (= Faktorwerte) untereinander.

In einer obliquen Faktorenanalyse werden die Fähigkeiten als beliebig korreliert gedacht, in der orthogonalen Faktoranalyse werden sie als unabhängig vorausgesetzt

bzw. definiert. Die Annahme unabhängiger Faktoren führt zu einigen mathematischen Vereinfachungen.

So ergibt sich in der orthogonalen Faktorenanalyse die Korrelation zwischen zwei Tests i und j allein aus den Ladungen in den gemeinsamen Faktoren, wie in [4.3] angegeben:

[4.3] $$r_{ij} = a_{i1}\, a_{j1} + a_{i2}\, a_{j2} + \ldots$$

Darüber hinaus läßt sich in der orthogonalen Faktorenanalyse die Ladung zugleich als die Korrelation des Tests mit den Faktorwerten in diesem Faktor interpretieren. Weiterhin läßt sich bei orthogonalen Faktoren die beobachtbare Testvarianz in additive Anteile aufspalten, die auf die einzelnen Faktoren zurückgehen (Gleichung [4.4]):

[4.4] $$\sigma^2(z_i) = a^2_{i1} + a^2_{i2} + \ldots. + \sigma^2(u_i)$$

Das Quadrat der Ladung (a^2) gibt somit den Anteil an der Testvarianz an, der auf den entsprechenden Faktor (auf individuelle Unterschiede in der entsprechenden Fähigkeit) zurückzuführen ist. Die Summe der Ladungsquadrate für einen Test heißt Kommunalität und gibt an, zu welchem Anteil die Varianz dieses Tests durch die gemeinsamen Faktoren "aufgeklärt" wird. Sie wird gewöhnlich mit h^2 bezeichnet. Zur globalen Charakterisierung, inwieweit in einer Faktoranalyse die Varianz aller Variablen durch die gemeinsamen Faktoren aufgeklärt wird, kann man die durchschnittliche Kommunalität angeben. Zur Charakterisierung dessen, wieviel jeder einzelne Faktor zur aufgeklärten Varianz aller Variablen beiträgt, kann man die Summe der Ladungsquadrate dieses Faktors (summiert über die Variablen) zur Summe der Kommunalitäten in Beziehung setzen.

4.2.1.2 Geometrische Darstellung, Rotationsproblem, Kommunalitätenproblem

Die rechnerische Aufgabe der Faktorenanalyse besteht darin, aus den Korrelationen aller Tests untereinander die Faktorladungen zu bestimmen. Dabei ist man bestrebt, mit möglichst wenigen Faktoren auszukommen und dabei die Faktorladungen so zu bestimmen, daß man aus ihnen die beobachteten Korrelationen möglichst genau reproduzieren kann. Hat man es z.B. mit 20 Tests zu tun, deren Korrelationen aus 3 Faktoren erklärt werden sollen, so müssen sich die 190 beobachteten Korrelationen aus nur 20 x 3 = 60 Faktorladungen jeweils gemäß Gleichung [4.3] ergeben.

Wie man rechnerisch vorgeht, um dies in bestmöglicher Näherung zu erreichen und wie man entscheidet, ob weitere Faktoren notwendig sind, kann hier nicht dargestellt werden. Statt dessen sollen hypothetische Ausgangsdaten und das Ergebnis einer orthogonalen Faktorenanalyse die Grundgleichungen an einem Zahlenbeispiel illustrieren. An diesem Beispiel soll dann auch die geometrische Darstellung und das Rotationsproblem erläutert werden.

Tabelle 4.1a enthält die Korrelationen zwischen 6 Tests (fingierte Daten), Tabelle 4.1b die Faktorladungen in zwei gemeinsamen Faktoren und Tabelle 4.lc die aus den Faktorladungen gemäß Gleichung (4.3) rekonstruierten Korrelationen. Die Abweichungen zwischen den Korrelationen in den Daten und den rekonstruierten Korrelationen (= Residuen) sind hier so gering, daß man zwei Faktoren als zur Erklärung der Korrelationen ausreichend ansehen wird (Tabelle 4.ld).

Tabelle 4.1a: Korrelationen zwischen 6 Tests (fingierte Daten)

	Test					
	1	2	3	4	5	6
1	-	.41	.43	.30	.33	.19
2	-	-	.70	.50	.56	.30
3	-	-	-	.54	.63	.36
4	-	-	-	-	.68	.39
5	-	-	-	-	-	.44
6	-	-	-	-	-	-

Tabelle 4.1b: Faktorladungen der 6 Tests in den 2 Faktoren und Kommunalitäten der Tests

		Faktoren		Kommunalität
		I	II	h^2
Tests	1	.46	.20	.21
	2	.74	.32	.55
	3	.83	.36	.69
	4	.35	.70	.61
	5	.44	.74	.74
	6	.25	.44	.26

Tabelle 4.1c: Aus den in Tabelle 4.1b angegebenen Faktorladungen gemäß Gleichung [4.3] rekonstruierte Korrelationen zwischen den 6 Tests

	Test					
	1	2	3	4	5	6
1	-	.40	.45	.30	.35	.20
2	-	-	.72	.48	.56	.32
3	-	-	-	.54	.63	.36
4	-	-	-	-	.67	.39
5	-	-	-	-	-	.43
6	-	-	-	-	-	-

Tabelle 4.1d: Residuen
Differenzen zwischen den Ausgangskorrelationen in Tabelle 4.la und den aus den Faktorladungen rekonstruierten Korrelationen in Tabelle 4.lc

	Test					
	1	2	3	4	5	6
1	-	.Ol	-.02	.00	-.02	-.0l
2	-	-	-.02	+.02	.00	-.02
3	-	-	-	.00	.00	.00
4	-	-	-		+.01	.00
5	-	-	-	-	-	.01
6	-	-	-	-	-	-

Geometrische Darstellung

Bei nur zwei Faktoren lassen sich die Ergebnisse einer Faktorenanalyse leicht graphisch veranschaulichen. In Abbildung 4.1 sind die Faktoren I und II als Achsen eines Koordinatensystems dargestellt und die Tests sind gemäß ihren Ladungen eingetragen.

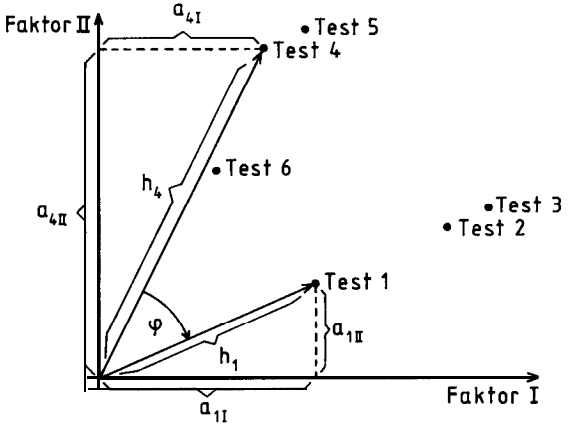


Abbildung 4.1: Darstellung von 6 Tests im zweidimensionalen Faktorraum. Ladungen gemäß Tabelle 4.1b

Es läßt sich zeigen (pythagoräischer Lehrsatz), daß die Wurzel aus der Kommunalität der Länge des Vektors eines Tests (graphisch als Pfeil vom Nullpunkt des Koordinatensystems zum Test hin dargestellt) entspricht. Weiter ergibt sich die Korrelation zweier Tests aus der Länge ihrer Vektoren und dem eingeschlossenen Winkel

(ableitbar aus dem Cosinus-Satz der Geometrie), wie in Gleichung [4.5] angegeben.

[4.5] $$r_{ij} = a_{i1}\, a_{j1} + a_{i2}\, a_{j2} = h_i\, h_j \cos \varphi$$

Diese Beziehungen gelten bei mehr als zwei Faktoren im mehrdimensionalen Raum entsprechend.

Das Rotationsproblem

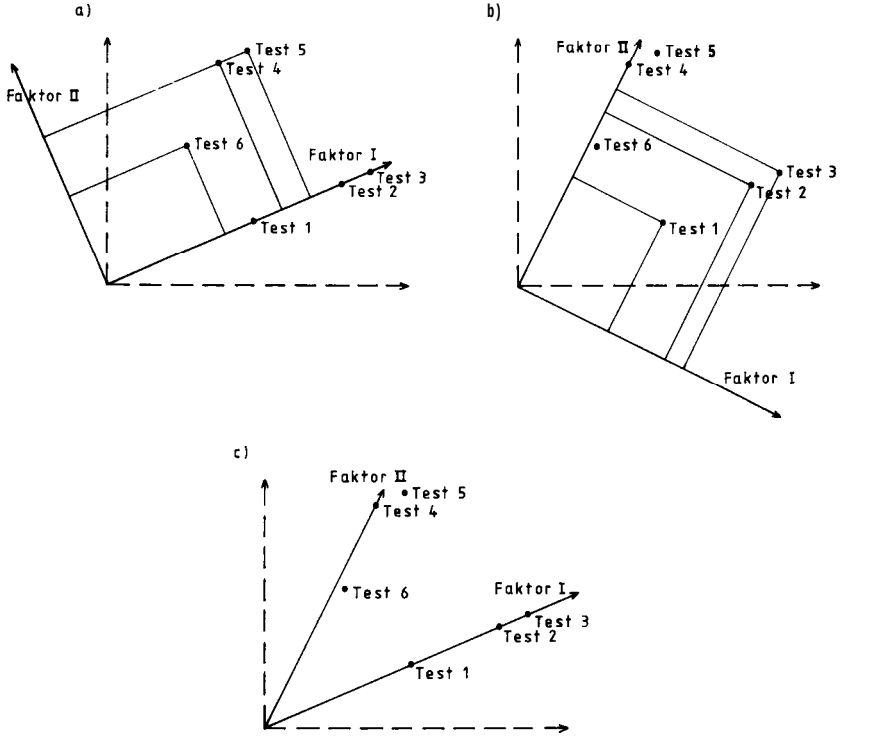
Wenn man eine Faktorlösung gefunden hat, so kann man dazu beliebig viele weitere konstruieren, die zu genau denselben rekonstruierten Korrelationen führen, also ganz genauso gut auf die Daten passen: Wie in Gleichung [4.5] angegeben, hängen die Korrelationen zwischen den Tests nur von der Länge ihrer Vektoren und den eingeschlossenen Winkeln ab. Wenn nun das Koordinatensystem beliebig gedreht wird, so ändert sich an diesen Winkeln und Längen nichts (d.h. die reproduzierten Korrelationen bleiben gleich), wohl aber an den Koordinaten der Tests (den Faktorladungen), die nun am neuen Koordinatensystem abzulesen sind. Indem man das Koordinatensystem beliebig dreht, kann man somit beliebig viele mathematisch gleichwertige Lösungen für die Faktorladungen produzieren. Diese können inhaltlich recht unterschiedliche Deutungen nahelegen: Bei der in Abbildung 4.1 bzw. Tabelle 4.1b dargestellten Lösung kommen die Korrelationen zwischen den Tests durch zwei Faktoren zustande, auf denen alle Tests Ladungen haben. Durch eine Drehung des Koordinatensystems um ca. 24 Grad nach links erhält man die in Abbildung 4.2a angegebene Lösung mit einem Generalfaktor und einem Gruppenfaktor. Die Tests 1,2,3 laden nur im Generalfaktor während die Tests 4, 5, 6 einen zusätzlichen Faktor gemeinsam haben. Durch eine Drehung des Koordinatensystems um ca. 26 Grad nach rechts entsteht das in Abbildung 4.2b dargestellte Bild: Nun scheinen die Tests 4,5 und 6 nahezu nur einen Generalfaktor zu erfassen, während die Tests 1,2,3 einen zusätzlichen Faktor gemeinsam haben. Das Problem, zwischen solchen mathematisch gleichwertigen, aber inhaltlich verschiedenen Lösungen zu entscheiden, ist als Rotationsproblem der Faktorenanalyse bekannt. Über mathematische Versuche zu definieren, was eine "einfache" und damit gut interpretierbare Lösung ist, und über die rechnerische Durchführung der Rotation bei mehr als zwei Faktoren soll hier nicht berichtet werden. Das Thema ist in den am Ende des Kapitels genannten Lehrbüchern behandelt.

Schiefwinkelige Rotation und Faktoren zweiter Ordnung

Bei den in Abbildung 4.1 und 4.2a,b dargestellten Lösungen stehen die als Koordinaten gezeichneten zwei Faktoren jeweils im rechten Winkel aufeinander. Dem entspricht die Annahme, daß die beiden Faktoren unkorreliert (= orthogonal) sind. Faktor I könnte z.B. Rechenfähigkeit, Faktor II Wortflüssigkeit sein. Die Orthogonalität bedeutet dann, daß die beiden Fähigkeiten in der Personenpopulation nicht korrelieren. Wenn es gelänge, zwei Tests zu konstruieren, von denen der eine ausschließlich Rechenfähigkeit mißt (also auf der Ordinate liegt) und der andere ausschließlich Wortflüssigkeit mißt (also auf der Abszisse liegt), so wäre der Winkel zwischen den Vektoren dieser beiden Tests 90 Grad und gemäß Formel [4.5] müßten auch die beobachteten Testwerte zu Null korrelieren.

Geht man von der Forderung unkorrelierter Faktoren ab, so kann man weitere Lösungen produzieren, indem man die Koordination verschiedene Winkel bilden läßt

Abbildung 4.2: Rotation. Orthogonale Lösungen (a) und (b) und eine nicht-orthogonale Lösung (c)



Gegenüber der in Abbildung 4.1 dargestellten Lösung ist das Koordinatensystem in (a) um 24 Grad nach links, in (b) um 26 Grad nach rechts gedreht. In (c) bilden die Koordinaten einen Winkel von 40 Grad und entsprechen einer Faktorkorrelation von 0.77.

und jeweils die Koordinaten der Punkte in diesem schiefwinkeligen Koordinatensystem berechnet. Eine solche Lösung ist in Abbildung 4.2c angegeben. Die Koordinaten bilden hier einen Winkel von 41 Grad, was einer Korrelation der Faktoren von cos (41°) = 0.76 entspricht. Sowohl die Tests 1, 2, 3 als auch 4,5,6 liegen fast genau auf einer Koordinatenachse. Die Tests messen also jeweils (fast) nur eine der beiden Fähigkeiten, die beiden Fähigkeiten sind aber miteinander korreliert. Auch dies wäre eine inhaltlich plausible Deutung des Korrelationsmusters.

Wenn man es nicht nur mit zwei, sondern mit mehreren Faktoren zu tun hat und sich für eine Lösung mit korrelierenden Faktoren entschieden hat, so kann man weiter fragen, wie denn die Korrelationen zwischen den Faktoren zustande kommen. Nimmt man die Korrelationen zwischen den Faktoren als "Daten" und unterzieht sie ihrerseits einer Faktoranalyse, so nennt man die daraus resultierenden Faktoren "Faktoren zweiter Ordnung", und es entsteht ein hierarchisches Modell: Die Testleistungen werden aus gewichteten Summen von Fähigkeiten erklärt, die Korrelationen zwischen den Fähigkeiten aus Faktoren zweiter Ordnung (z.B. Fähigkeiten wie Wort-

schatz, Worteinfallgeschwindigkeit, Erkennen verbaler Beziehungen aus einem allgemeineren verbalen Faktor und spezifischeren Komponenten). Auch Faktorenanalysen zweiter Ordnung lassen ihrerseits wieder Spielraum für Rotation.

Aufgrund der dargestellten Vielfalt äquivalenter Modelle wird man wohl kaum den Anspruch erheben, mittels Faktorenanalyse eine bestimmte Lösung als die richtige ausweisen zu können. So z.B. lassen die positiven Korrelationen zwischen Intelligenztests, die man in aller Regel findet (selbst wenn sich der Testautor um Unabhängigkeit der Einzeltests bemüht hat, z.B. Intelligenz-Struktur-Test von Amthauer, 1953, 1970; Leistungsprüfsystem von Horn, 1962), verschiedene inhaltliche Deutungen zu:

(a) Es gibt einen Faktor der allgemeinen Intelligenz, der in alle Testleistungen mehr oder weniger stark eingeht.

(b) Es gibt keinen allgemeinen Faktor, sondern Intelligenz besteht aus mehreren unabhängigen Einzelfähigkeiten. Es ist aber nicht möglich, faktoriell reine Tests zu konstruieren, sondern jeder Test beansprucht mehrere Fähigkeiten.

(c) Es gibt keinen allgemeinen Faktor, sondern mehrere Einzelfähigkeiten. Die Tests (oder Testgruppen) erfassen jeweils nur eine dieser Fähigkeiten, die Fähigkeiten sind miteinander korreliert.

Jede dieser inhaltlichen Deutungen läßt sich in ein faktorenanalytisches Modell umsetzen, jedes ist mit den Daten vereinbar. Diese Unentscheidbarkeit, die als Rotationsproblem schon im Modell-Ansatz enthalten ist, ist ein Grund dafür, daß man von einer Faktoranalyse keine abschließenden Aussagen über die einer Testleistung zugrundeliegenden Funktionen und Prozesse erwarten kann.

Praktisch bevorzugt werden möglichst einfache, gut interpretierbare Lösungen. Bei den meisten Arbeiten, die Faktorenanalysen anwenden, werden orthogonale Lösungen gewählt, und es wird zu einem mathematisch definierten Einfachheitskriterium rotiert. Das bekannteste ist das Varimax-Kriterium: Pro Faktor wird die Varianz der quadrierten Ladungen berechnet; die Summe dieser Varianzen ist das Kriterium, das maximiert wird. Die Varianz der quadrierten Ladungen wird groß, wenn sowohl Null-Ladungen als auch dem Betrag nach hohe Ladungen vorhanden sind. Ein in diesem Sinn prägnantes Ladungsmuster läßt sich im allgemeinen leichter inhaltlich deuten als ein Ladungsmuster mit vielen mittleren Ladungen auf allen Variablen.

Das Problem der Kommunalitätenschätzung

Die rechnerische Durchführung der Faktorenanalyse setzt nicht nur die Kenntnis der Korrelationen, sondern auch die Kenntnis der Kommunalitäten (zum Begriff der Kommunalität siehe Abschnitt 4.2.1.1) voraus. Diese aber ergeben sich erst aus den zunächst noch unbekannten Faktorladungen. Es gibt zwar verschiedene Möglichkeiten, die Kommunalitäten schon vorher zu schätzen, doch kann das Ergebnis einer Faktorenanalyse nicht nur bezüglich der Höhe der Ladungen, sondern auch bezüglich der Zahl der benötigten Faktoren von der Wahl des Kommunalitäten-Schätzverfahrens abhängen. Dieses Problem, das ebenfalls schon im mathematischen Ansatz der Faktorenanalyse steckt, trägt zur weiteren Uneindeutigkeit faktorenanalytischer Lösungen bei.

Die Hauptkomponenten-Analyse (englisch: principal component analysis) unterscheidet sich vom klassischen Ansatz der Faktorenanalyse, wie er in Formel [4.2] angegeben ist, dadurch, daß keine Uniqueness vorgesehen ist und alle Kommunalitäten

gleich Eins sind. Da jeder Test Meßfehler enthält, ist dieser Ansatz mit einer funktionalen Interpretation von vornherein nicht vereinbar. Ziel ist hier lediglich, die Vielzahl von Testvariablen auf einige wenige Faktoren zu reduzieren, die die in den Testvariablen enthaltene Information möglichst gut repräsentieren. Näheres zur Beziehung zwischen klassischer Faktoranalyse und Hauptkomponentenanalyse findet man bei Snook & Garsuch (1989) und Velicer & Jackson (1990).

4.2.2 Haupteinwände gegen die Faktorenanalyse als erklärende Theorie

Bereits im vorangehenden Kapitel wurde deutlich, daß die Ergebnisse der Faktorenanalyse mathematisch nicht eindeutig sind, sondern dem Forscher einen erheblichen Interpretationsspielraum lassen. Das betrifft sowohl die Anzahl der Faktoren, die je nach dem gewählten Kommunalitäten-Schätzverfahren und je nach Abbruchkriterium für die Faktorenextraktion unterschiedlich ausfallen kann, als auch die Festlegung der Rotation. Allein diese Unbestimmtheit mag die Faktorenanalyse als "weiche" Methode erscheinen lassen, wenig geeignet für eine stringente Überprüfung von Theorien.

Die Haupteinwände dagegen, daß man mittels Faktorenanalysen die Grundfähigkeiten entdecken und das Zustandekommen von Testleistungen erklären, also die Grundgleichung allgemeinpsychologisch auffassen und funktional interpretieren könnte, sind jedoch nicht nur in dieser mathematischen Unterbestimmtheit zu sehen, sondern vor allem in einer Reihe von Kritikpunkten, die Ende der Sechzigerjahre von verschiedener Seite (Fischer, 1968; 1974; Kallina, 1967; Kalveram, 1965; 1970a, b; Merz & Kalveram, 1965) vorgetragen wurden: Die prinzipielle Unüberprüfbarkeit des Ansatzes, die Populationsabhängigkeit der Ergebnisse, die Entstehung von Artefakten durch simultane Überlagerung oder Selektionseffekte.

Zur Unüberprüfbarkeit des Ansatzes

In der Grundgleichung (Gleichung [4.2]) wird angenommen, daß die Testleistung aufgrund einer gewichteten Summe von Fähigkeiten zustande kommt, wobei
(a) die Gewichtung für alle Personen gleich ist und
(b) die Fähigkeiten einander beliebig kompensieren können.

Als Ausgangsdaten für eine Faktorenanalyse stehen aber nur die Korrelationen zwischen den Tests zur Verfügung. Egal wie diese zustande gekommen sind - ob gemäß den in der Grundgleichung ausgedrückten Annahmen oder ganz anders -,jede Korrelationsmatrix kann faktorisiert werden, und es ist dem Ergebnis der Faktorenanalyse nicht anzusehen, ob die Annahmen der Grundgleichung zutreffen oder nicht.

Die Populationsabhängigkeit des Ergebnisses

Korrelationen beschreiben Merkmalszusammenhänge in Populationen. Sie können in unterschiedlichen Populationen (definiert nach Alter, Geschlecht, Schulbildung usw.) unterschiedlich ausfallen. Dementsprechend wird auch das Ergebnis einer Faktorenanalyse derselben Tests, sowohl was die Anzahl der Faktoren als auch was die Ladungen anbelangt, von Population zu Population unterschiedlich sein. Andererseits gehört aber jede einzelne Person mehreren Populationen zugleich an: eine 13jährige

Oberschülerin z.B. der Population der 13jährigen, der Population der Mädchen, der Population der Oberschülerinnen. Interpretiert man das Ergebnis einer Faktorenanalyse auf individueller Ebene als Aussage darüber, wieviele und welche Fähigkeiten eine Person für die Lösung des Tests einsetzt, so gerät man sehr bald in Widersprüche. Derselben Person wäre je nachdem, welcher Population man sie gerade zurechnet, eine andere Fähigkeitsstruktur zuzuschreiben.

Artefakte durch simultane Überlagerung und Selektionseffekte

Selbst wenn die Testleistung in einer Population bei jedem Einzelindividuum so zustande kommt, wie in der Grundgleichung angenommen, ist nicht gewährleistet, daß man als Ergebnis der Faktorenanalyse die richtige Zahl von Faktoren und richtigen Ladungen erhält. Das haben Merz & Kalveram (1965) am Beispiel der Differenzierungshypothese der Intelligenz eindrucksvoll gezeigt:

Gemäß der Differenzierungshypothese ändert sich die Intelligenz in der Entwicklung vom älteren Kind **zum** Erwachsenen vor allem qualitativ durch Differenzierung. Dementsprechend wird mit dem Alter ein Absinken der Korrelationen zwischen den Tests, eine Zunahme der Zahl unabhängiger Fähigkeiten und eine Abnahme der Bedeutung des Generalfaktors erwartet. Merz & Kalveram (1965) konnten zeigen, daß dasselbe Ergebnis zu erwarten ist, wenn die Intelligenzstruktur, was Anzahl und Gewicht der zur Lösung eingesetzten Faktoren anbelangt, gleichbleibt, auf den einzelnen Altersstufen aber unterschiedlich starke individuelle Differenzen im allgemeinen Entwicklungsstand bestehen. Besonders auf den unteren Altersstufen, wo **das** Entwicklungstempo noch rasch ist, werden manche Kinder gegenüber den Gleichaltrigen einen alle Fähigkeiten mehr oder weniger stark betreffenden Entwicklungsvorsprung, andere einen Entwicklungsrückstand haben. Wenn alle Testleistungen eines Probanden zugleich (simultan) in dieselbe Richtung beeinflußt werden, steigen die Korrelationen zwischen den Tests. Merz & Kalveram (1965) sprechen von "simultaner Überlagerung" der Korrelationsstruktur durch Kovarianz, die auf Unterschiede im Entwicklungsstand zurückgeht. Im Erwachsenenalter da**gegen,** wenn die Entwicklung praktisch abgeschlossen ist, spielen diese Entwicklungsunterschiede keine Rolle mehr, und die Korrelationen fallen niedriger aus. Als Ergebnis von orthogonalen Faktoranalysen erhält man bei den Jüngeren höhere Kommunalitäten, einen stärkeren Generalfaktor, geringere Ladungen in den weiteren Faktoren und -je nach Abbruchkriterium - eine geringere Gesamtzahl von Faktoren. Insgesamt entsteht also ein Bild, das voll den Erwartungen aufgrund der Differenzierungshypothese gleicht. Weitere Beispiele für Artefakte durch simultane Überlagerung sind in derselben Arbeit und bei Kalveram (1965) zu finden.

Eine weitere Quelle von Artefakten, die die Korrelationen zwischen den Tests so verändern können, daß selbst dann, wenn die Grundgleichung als Annahme über den Lösungsprozeß bei jedem einzelnen Probanden zutrifft, die Faktorenanalyse als Ergebnis weder die richtige Faktorenzahl noch die richtigen Ladungen liefert, sind Selektionseffekte. Kalveram (1969) demonstriert an einem Beispiel mit Intelligenztestdaten, daß schon eine mäßige Selektion nach der Punktsumme (Weglassen der Probanden mit den höchsten und niedrigsten Werten für den Gesamt-IQ) deutliche Effekte auf die lnterkorrelationen der Tests hat: Extreme Summenwerte kommen zustande, wenn Probanden in allen Tests gut oder in allen Tests schlecht abgeschnitten haben. Ein Weglassen dieser Fälle muß zu einer Reduktion der Korrelationen führen. In einem so selegierten Datenmaterial sind dann weder die gemeinsamen Faktoren voneinander unabhängig, wie das in der orthogonalen Faktorenanalyse vorausgesetzt wird, noch **auch** die spezifischen von dem gemeinsamen (eine Voraus-

setzung, die auch in der obliquen Faktorenanalyse gemacht wird), und das Ergebnis der Faktorenanalyse wird in die Irre führen.

Dem kann man nun entgegenhalten, daß eine explizite Selektion an den Daten ja in der Regel nicht erfolgt. Andererseits kann auch in "natürlichen" Populationen, wie z.B. Schülern einer bestimmten Schulart mit mittlerem Anforderungsniveau, faktisch eine Selektion nach dem Durchschnittsniveau eines Schülers (Mittel über seine Fähigkeiten) stattgefunden hat, indem Extremfälle positiver wie negativer Art die Schule häufiger verlassen haben. In diesem Falle gelten die obigen Argumente entsprechend.

Aufgrund der genannten Argumente wurde der Anspruch aufgegeben, die Ergebnisse von Faktorenanalysen könnten als für jeden einzelnen Probanden gültige Aussage über das Zustandekommen von Testleistungen interpretiert werden.

Ungeachtet dessen bleibt das Problem bestehen, daß man bei der Beurteilung der Validität eines Tests weitgehend auf Korrelationen angewiesen ist und größere Mengen von Korrelationen konsistent interpretieren möchte. Da die Faktorenanalyse solche Interpretationsmöglichkeiten aufzeigen kann, wurde sie trotz des Vorwurfs der Nicht-Falsifizierbarkeit als Theorie, als heuristisches Instrument auch in Zeiten stärkster Kritik unvermindert zum Einsatz gebracht. Sie wird dann als eine datenexplorierende Technik aufgefaßt, die mit dem Korrelationsmuster vereinbare Deutungen anbietet, wobei man freilich zunächst nicht weiß, ob eine davon richtig ist und welche. Die Entscheidung darüber, welche Hypothesen weiter verfolgt werden sollen, ist dann nur aufgrund zusätzlicher Information aus inhaltlichen Gründen möglich.

Eine noch entschiedenere Abkehr vom ursprünglichen Anspruch vollzieht man, wenn man die Faktoranalyse als eine Methode auffaßt, die für eine bestimmte Population Dimensionen individueller Unterschiede beschreibt. Da Beschreibungsdimensionen nur nach Gesichtspunkten der Zweckmäßigkeit, Ökonomie und Ergiebigkeit zu beurteilen sind, nicht aber nach "wahr" oder "falsch", stellt sich die Frage nach der Falsifizierbarkeit erst gar nicht. Daß bei Populationen, die sich in Art und Ausmaß individueller Unterschiede unterscheiden, jeweils andere Beschreibungsdimensionen in den Vordergrund treten, erscheint dann als selbstverständlich und sachlich begründet und nicht als Mangel der Methode. Auch das Problem der Artefakte, z.B. durch simultane Überlagerung oder Selektionseffekte, stellt sich erst, wenn man über die Definition von Beschreibungsdimensionen hinausgeht und nach den Gründen fragt, warum z.B. in der einen Altersklasse ein Generalfaktor den größten Teil der Varianz abschöpft, in der anderen nicht. Die Beschreibung des Sachverhalts läßt mehrere Deutungen (Differenzierungshypothese, simultane Überlagerung) zu, zwischen denen erst durch zusätzliches Wissen (hier über Entwicklungskurven und das Ausmaß individueller Unterschiede auf den einzelnen Altersstufen) zu entscheiden ist.

Eine typische Anwendung dieser Art, bei der die Faktorenanalyse von vornherein nur mit dem Ziel eingesetzt wird, eine Vielzahl von Variablen auf eine oder einige wenige Beschreibungsdimensionen zu reduzieren, die die wesentliche Information enthalten, liegt z.B. vor, wenn aus einer Vielzahl von Intelligenztests ein Gesamtwert gebildet werden soll, der dann in der weiteren Auswertung anstelle der vielen Einzeltests die Intelligenz repräsentieren soll. Hier liegt es nahe, aus einer Faktorenanalyse nach der Hauptkomponentenmethode die erste Hauptkomponente (den Faktor, der die meiste Varianz abschöpft) zu verwenden. Eine weitere Anwendung, die mit einer Deutung der Faktorenanalyse als Methode zur bloß deskriptiven Dimensionsanalyse auskommt, ist die Faktorenanalyse von Testitems mit dem Ziel, ltemgruppen zu Skalen zusammenzustellen, die für diese Population (!) eine hohe innere Konsistenz der

Skalen erwarten lassen. Auch bei völliger Rücknahme des Anspruchs auf ein bloßes Datenreduktionsverfahren lassen sich also sinnvolle Anwendungen für die Faktorenanalyse finden.

Eine wesentliche Weiterentwicklung der klassischen Faktorenanalyse, die inzwischen oft auch als "exploratorische" Faktorenanalyse bezeichnet wird, stellt die konfirmatorische Faktorenanalyse dar. Sie geht von inhaltlichen Hypothesen aus und macht falsifizierbare Aussagen über die Struktur der Korrelations- oder Kovarianzmatrix. Einige Einsatzmöglichkeiten im Rahmen der Testtheorie sollen im folgenden an Beispielen dargestellt werden. Dabei wird sich freilich auch zeigen, daß auch eine falsifizierbare Theorie, wenn sie auf die Daten paßt, deshalb noch lange nicht die einzige mögliche Erklärung sein braucht. Wenn die Vorhersagen der Theorie sehr strikt sind, wird es allerdings sehr schwer werden, plausible Alternativerklärungen für dieselben Daten zu finden.

4.2.3 Einsatzmöglichkeiten und Grenzen der konfirmatorischen Faktorenanalyse

In der klassischen Faktorenanalyse braucht der Forscher kein Vorwissen über die Anzahl der Faktoren oder über das zu erwartende Ladungsmuster zu besitzen. Es werden so lange Faktoren extrahiert, bis die Korrelationsmatrix aus den Faktorladungen hinreichend genau reproduzierbar ist. Es wird dann durch Rotation (nach mathematischen oder inhaltlichen Kriterien) eine gut interpretierbare Lösung gesucht. In der konfirmatorischen Faktorenanalyse, die - ausgehend von den Arbeiten von Jöreskog (1967; 1969) -vor allem in den siebziger Jahren entwickelt wurde, muß der Forscher schon vor Eintritt in das Verfahren eine Hypothese über die Zahl der Faktoren und das Ladungsmuster haben. Bei der Hypothese über das Ladungsmuster handelt es sich in der Regel um Annahmen darüber, daß einzelne Tests auf bestimmten Faktoren nicht laden (vorgeschriebene Null-Ladungen), oder um Annahmen über Gleichheit bestimmter Ladungen (Gleichheits-Restriktionen). Darüber hinaus können über die Korrelationen der Faktorwerte einschränkende Annahmen gemacht werden (z.B. daß alle oder auch nur bestimmte Faktoren unkorreliert sind) und bezüglich der Residuen (testspezifische Faktoren und Meßfehler) Festlegungen getroffen werden (z.B. Gleichheit der Residualvarianzen bei bestimmten Tests). Insgesamt müssen die gesetzten Restriktionen ausreichen, um die Lösung mathematisch eindeutig zu machen, insbesondere also auch die Rotation festzulegen.

Ausgangsdaten können Korrelations- oder Kovarianzmatrizen sein. Die Parameter des Modells (Faktorladungen, Korrelationen der Faktoren, Residualvarianzen) werden dann so geschätzt, daß sie (a) den durch die Hypothese gesetzten Restriktionen genügen und (b) die empirischen Korrelationen (oder Kovarianzen) zwischen den Tests so gut, wie unter den gesetzten Restriktionen möglich, reproduzieren. Anhand der erreichten Anpassung (Übereinstimmung der aus den Ladungen reproduzierten Korrelationen mit den aus den Daten errechneten) wird beurteilt, ob die Hypothese mit den Daten vereinbar ist oder nicht.

Zur Schätzung der Parameter und zur Beurteilung der Anpassung stehen eine Reihe theoretisch unterschiedlich begründeter Verfahren zur Verfügung (eine neuere Übersicht findet man bei Anderson & Gerbing, 1988). Das am stärksten verbreitete Computer-Programm dürfte nach wie vor das Programm LISREL (zur Zeit neueste Version: LISREL 7, Jöreskog & Sörbom, 1989) sein, an zweiter Stelle dürfte das Pro-

gramm EQS (Bentler, 1985) stehen. Beide Programme umfassen einen weiten Bereich von linearen Strukturgleichungsmodellen und enthalten die konfirmatorische Faktorenanalyse als Spezialfall.

Im folgenden soll an vier unterschiedlichen Fragestellungen gezeigt werden, wie Problemstellungen aus der Testtheorie mit Hilfe konfirmatorischer Faktorenanalysen bearbeitet werden können.

Beispiel 1 : Überprüfung der Parallelität von Tests

Zwei oder mehr Tests sind parallel im Sinne der klassischen Testtheorie, wenn sie dieselben wahren Werte und gleiche Meßfehlervarianzen haben. Daraus folgt u.a., daß ihre Varianzen gleich sind, daß die Kovarianzen der Parallelformen untereinander gleich sind und daß die Kovarianzen der Parallelformen zu einem beliebigen Außenkriterium gleich sind. Diese Struktur der Kovarianzmatrix kann in einer konfirmatorischen Faktorenanalyse überprüft werden. Die Parallelitätshypothese wird dabei ausgedrückt, indem für die Tests festgelegt wird, (a) daß sie auf einem gemeinsamen Faktor laden, (b) daß die Ladungen auf diesem Faktor gleich sind und (c) daß die Residualvarianzen gleich sind. Abbildung 4.3 zeigt ein hypothetisches Beispiel:

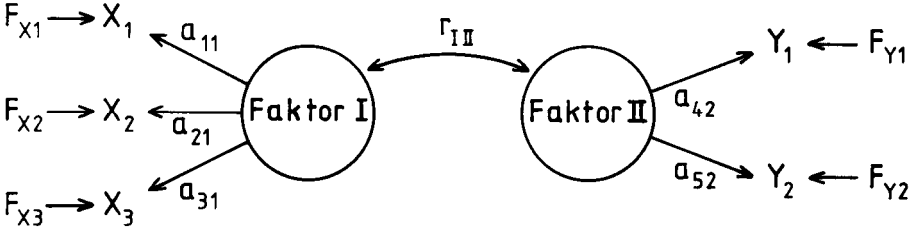


Abbildung 4.3: Konfirmatorische Faktoranalyse zur Prüfung der Parallelität der Tests X_I, X_2, X_3 und Y_1, Y_2. Parameterspezifikation und Restriktionen s. Tabelle 4.2

Die Tests X_1, X_2, X_3 sollen drei Parallelformen eines Wortschatztests sein, die Tests Y_1 und Y_2 zwei Parallelformen eines Rechentests. Der Anteil, den Faktor I an der Varianz eines Wortschatztests ausmacht, entspricht der wahren Varianz; die Restvarianz ist die Fehlervarianz. Entsprechendes gilt für Faktor II und die beiden Rechentests. Die Korrelation der Faktoren I und II ist die Korrelation der wahren Werte von Wortschatztest und Rechentest. Will man das Modell prüfen, so hat man die Parametermatrizen zu spezifizieren und die Restriktionen zu setzen, wie in Tabelle 4.2 angegeben.

Wenn das Modell nicht paßt, kann eine schwächere Hypothese geprüft werden: Beispielsweise könnten die Tests X_1, X_2, X_3 dasselbe messen, aber mit unterschiedlicher Reliabilität. Will man unterschiedliche wahre Varianzen zulassen, so ist die Gleichheitsrestriktion für Faktor I im Ladungsmuster aufzuheben; will man zusätzlich unterschiedliche Fehlervarianzen zulassen, so entfällt die entsprechende Restriktion bezüglich der Residualvarianzen.

Beispiele, in denen verschieden streng gefaßte Modelle an realen Daten (Intelligenz- und Schulleistungstest) vergleichend geprüft wurden, findet man in der inzwischen als klassisch anzusehenden Arbeit von Jöreskog (1978).

Tabelle 4.2: Parameterspezifikation zu dem in Abbildung 4.3 dargestellten Modell einer konfirmatorischen Faktoranalyse zur Prüfung der Parallelität der Tests X_1, X_2, X_3 und Y_1, Y_2.

	Ladungsmatrix Faktoren	
	I	II
Tests		
X_1	a_{11}	0
X_2	a_{21}	0
X_3	a_{31}	0
Y_1	0	a_{42}
Y_2	0	a_{52}

	Kovarianzmatrix der Faktoren	
	I	II
II	1	
II	$r_{1.11}$	1

Gleichheitsrestriktionen für die
(a) Ladungen:
$a_{11} = a_{21} = a_{31}$
$a_{42} = a_{52}$

(b) Fehlervarianzen:
$\sigma^2(F_{X1}) = \sigma^2(F_{X2}) = \sigma^2(F_{X3})$
$\sigma^2(F_{Y1}) = \sigma^2(F_{Y2})$

Beispiel 2: Überprüfung von Hypothesen über die Gleichheit von Ladungsmustern in verschiedenen Populationen
Wenn man die Anwendungen der klassischen Faktorenanalyse überblickt, so findet man ganz überwiegend orthogonale Faktorenlösungen. Werden analoge Analysen für unterschiedliche Personenstichproben durchgeführt, so wird in der Regel jede Faktorenanalyse für sich gerechnet und die Ergebnisse hinterher vergleichend diskutiert. Methoden zur Ähnlichkeitsrotation wurden zwar vorgeschlagen (z.B. Fischer & Roppert, 1964), aber kaum angewendet.

Andererseits ist es alles andere als plausibel anzunehmen, die Faktoren seien in den unterschiedlichsten Populationen immer wieder unkorreliert. Wenn die Faktoren in den Populationen unterschiedlich korreliert sind, wird eine von der Methode her gesetzte Orthogonalitätsrestriktion zu von Population zu Population unterschiedlichen Faktorlösungen führen - auch dann, wenn die Tests in allen Populationen dasselbe messen.

Ein Vorteil des Programms LISREL (Jöreskog & Sörbom, 1989) besteht darin, daß es die Möglichkeit bietet, an mehrere Datensätze simultan eine konfirmatorische Faktorenanalyse anzupassen. Dabei kann festgelegt werden, daß bestimmte Parameter (z.B. Faktorladungen) für alle Datensätze gleich sein sollen, während andere (z.B. Varianzen und Kovarianzen der Faktoren) von Stichprobe zu Stichprobe variieren können. Man kann also z.B. der Reihe nach folgende, zunehmend restriktive Modelle testen:

1. Dasselbe Ladungsmuster (Zuordnung der Tests zu den Faktoren und entsprechend vorgeschriebene Null-Ladungen) paßt in allen Populationen. Die Ladungen können aber in den einzelnen Populationen unterschiedlich hoch sein und die Faktoren können in den einzelnen Populationen unterschiedlich korreliert sein.
2. Ladungsmuster und Ladungen müssen in allen Populationen übereinstimmen, Faktorvarianzen und Faktorkorrelationen können aber von Population zu Population unterschiedlich sein.
3. Die Lösungen stimmen völlig überein, d.h. die Korrelationsmatrix ist in allen Populationen gleich.

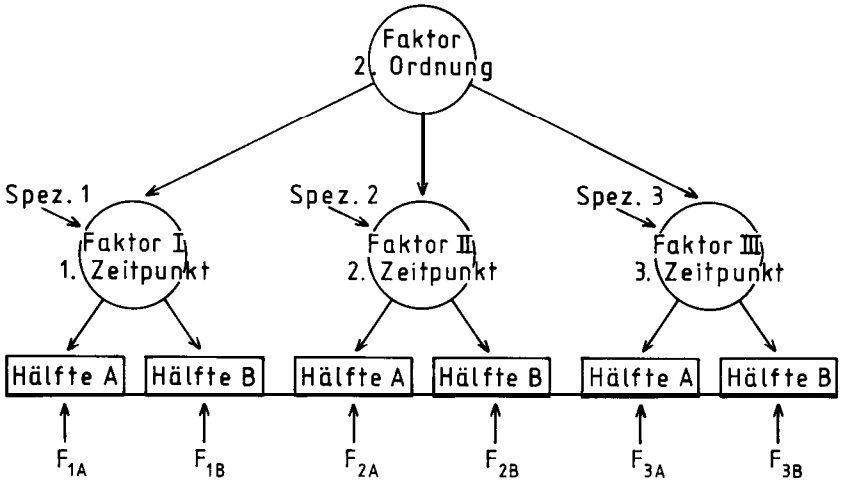
Ein Beispiel für eine solche schrittweise Anpassung einer gemeinsamen faktoranalytischen Lösung für 5 Datensätze findet man bei Schmidt (1983). Er untersuchte an 5 Altersgruppen die faktorielle Struktur eines Fragebogens über Arbeitsorientierungen. Angenommen wurde schließlich ein Modell mit 4 Faktoren und derselben Zuordnung der Items zu den Faktoren (gleiches Ladungsmuster), wobei aber die Höhe der Ladungen in den einzelnen Altersgruppen unterschiedlich war. Ein komplexeres Beispiel, bei dem 6 Modellvarianten verglichen wurden, findet man bei Jöreskog & Sörbom (1985, Kapitel V 3). Stelzl (1987) illustriert an einem Beispiel mit hypothetischen Daten die Vorteile einer simultan konfirmatorischen Faktorenanalyse über alle Datensätze gegenüber getrennten klassischen Faktorenanalysen mit Rotation zur orthogonalen oder auch nicht-orthogonalen Einfachstruktur.

Beispiel 3: Die Zerlegung der wahren Varianz in Konsistenz und Spezifität nach Steyer (1987)

Steyer (1987) und Majcen, Steyer & Schwenkmezger (1988) schlagen eine konfirmatorische Faktorenanalyse zweiter Ordnung vor, um "Spezifität" und "Konsistenz" als Anteile der wahren Varianz eines Tests zu unterscheiden. Dazu wird der Test in zwei Hälften geteilt und beide Hälften den Probanden zu mehreren "Meßgelegenheiten", z.B. Zeitpunkten im Abstand von jeweils mehreren Wochen, vorgelegt.

Die Kovarianzmatrix der Daten wird dann nach dem in Abbildung 4.4 dargestellten Modell analysiert. Den beiden Testhälften zum selben Zeitpunkt wird jeweils ein gemeinsamer Faktor erster Ordnung unterstellt. Der Varianzanteil dieses Faktors an der Testvarianz ist die wahre Varianz des Tests. Den Kovarianzen zwischen den Zeitpunkten wird dann ein Generalfaktor als Faktor zweiter Ordnung unterstellt. Den Varianzanteil eines Tests, der durch diesen Generalfaktor erklärt wird, nennt Steyer "Konsistenz", den Anteil an der wahren Varianz, der nicht durch den Generalfaktor erklärt wird, "Spezifität". Den Generalfaktor interpretiert er als "Personfaktor" oder "Trait", die Spezifität als "Situations-" oder "Person-Situations-Interaktionsvarianz". Zum selben Meßzeitpunkt können sich die einzelnen Personen in unterschiedlichen Situationen befinden, z.B. ausgeschlafen oder verkatert sein (Situationsvarianz) und auf diese Situationen personspezifisch reagieren (Person-Situations-Interaktionsvarianz).

Abbildung 4.4: Konfirmatorische Faktorenanalyse zweiter Ordnung zur Unterscheidung von Konsistenz und Spezifität im Sinn von Steyer (1987)



Ein Test, bestehend aus den Hälften A und B, wird zu drei Zeitpunkten vorgegeben. Seine wahre Varianz besteht aus einem Anteil, der auf den Generalfaktor zweiter Ordnung zurückgeht, und einem Anteil, der für den jeweiligen Zeitpunkt spezifisch ist.

Beispiel 4: Die Multitrait-Multimethod-Matrix

Campbell & Fiske (1959) zeigten, wie man anhand einer sogenannten "Multitrait-Multimethod-Matrix" konvergente und diskriminante Validität psychologischer Messungen überprüfen kann: Dazu müssen mehrere Eigenschaften (= traits), z.B. Popularität und Expansivität eines Schülers, mit mehreren Methoden (Selbstauskunft, Rating durch andere, Verhalten in einer Gruppensituation, Rollenspiel) erfaßt worden sein und die Korrelationen aller Messungen (hier: 2 Eigenschaften und 4 Methoden = 8 Messungen) vorliegen. Diese Korrelationsmatrix heißt Multitrait-Multimethod-Matrix und soll eine bestimmte Struktur aufweisen:

Auch wenn man bei psychologischen Maßen immer davon ausgehen muß, daß nur ein Teil der Varianz auf die zu messende Eigenschaft zurückgeht und ein Teil methodenspezifisch ist, so wird man von einem guten Maß doch verlangen, daß der methodenspezifische Anteil gering ist. Dementsprechend sollten die Korrelationen zwischen Maßen, die dieselbe Eigenschaft mit unterschiedlichen Methoden erfassen, deutlich höher ausfallen (konvergente Validität) als die Korrelationen zwischen Maßen, die dieselbe Methode verwenden, aber unterschiedliche Eigenschaften erfassen (niedrige Korrelationen unterschiedlicher Eigenschaften = diskriminante Validität).

Bei mehr als zwei Eigenschaften kann man auch das Muster der Korrelationen der Eigenschaften untereinander betrachten. Wenn die Eigenschaften mit derselben Methode erfaßt wurden (alle durch Selbstauskunft oder alle durch Fremdbeurteilung), sollte sich jeweils dasselbe Korrelationsmuster ergeben, egal um welche Methode es sich handelt.

Die von Campbell & Fiske (1959) angestellten Überlegungen, auf die eine längere Diskussion folgte (dargestellt bei Schmitt et al., 1977) lassen sich gut in eine kon-

firmatorische Faktorenanalyse übertragen. Jeder Eigenschaft und jeder Methode entspricht ein Faktor, wobei jedes Maß hier auf einem Eigenschafts- und einem Methodenfaktor lädt. Diese Ladungen sollen bestimmt werden, alle anderen sind Null. Die Eigenschaftsfaktoren können untereinander korreliert sein, sollen aber von den Methodenfaktoren unabhängig sein. Beispiele solcher Anwendungen findet man u.a. bei Kenny (1976) Schmitt, Coyle & Saari (1977) Schwarzer (1983), Ostendorf et al. (1986). Eines der von Schwarzer (1983) vorgestellten Beispiele wird im folgenden Abschnitt dargestellt und diskutiert.

Grenzen der konfirmatorischen Faktorenanalyse:

Der Haupteinwand gegen die klassische Faktorenanalyse als Mittel zur Prüfung von Theorien über das Zustandekommen von Testleistungen liegt in der Nicht-Falsifizierbarkeit des theoretischen Ansatzes: Das Rechenverfahren führt immer zu einer Faktorlösung - auch dann, wenn die Testleistungen in Wirklichkeit ganz anders zustande kommen.

Die konfirmatorische Faktorenanalyse bringt in diesem Punkt eine Verbesserung: Wenn die Hypothese über das Ladungsmuster sehr restriktiv ist, so muß die Korrelationsmatrix eine ziemlich genau festgelegte Struktur haben, um mit der Hypothese vereinbar zu sein. Hat sie diese Struktur nicht, so wird das Modell verworfen.

Trotzdem bleiben grundlegende Probleme bestehen. Wenn das Modell verworfen wird, weil es nicht auf die Daten paßt, so kann das daran liegen, daß grundlegende Annahmen falsch sind. Es kann aber auch daran liegen, daß die Korrelationen z.B. durch Selektionseffekte (vgl. Kapitel 4.2.2) verzerrt sind. Wenn es z.B. durch eine Selektion nach der Summe aller Faktorwerte zu negativen Korrelationen auch der Residuen kommt, so wird ein ansonsten richtiges Modell mit unabhängigen Residuen verworfen (korrelierende Residuen sind bei konfirmatorischen Faktorenanalysen zwar nicht grundsätzlich unzulässig, führen aber sehr bald zu trivialen oder auch zu nicht identifizierbaren, d.i. nicht schätzbaren Modellen).

Wenn ein Modell nicht paßt, wird es meist nicht pauschal verworfen, sondern man sucht nach Korrekturmöglichkeiten und modifiziert einige weniger wichtige Annahmen, bis der Modelltest keine signifikanten Abweichungen mehr ausweist. Dabei läuft man Gefahr, das Modell im nachhinein an zufällige Eigenschaften der Stichprobe anzupassen. Eine Signifikanzprüfung des modifizierten Modells erfordert in jedem Fall einen neuen, unabhängigen Datensatz. Wenn auf der anderen Seite ein Modell gut an die Daten angepaßt ist, so schließt das nicht aus, daß andere, ebenfalls plausible Modelle ebensogut auf dieselben Daten passen. Auch zu einem hoch restriktiven Modell wie einer Multitrait-Multimethod-Lösung kann es Alternativen geben. Das soll durch eine Reanalyse eines von Schwarzer (1983) vorgestellten Beispiels demonstriert werden. Schwarzer (1983) verwendete die Daten von Winne & Marx (1981, zit. nach Schwarzer, 1983), um eine Multitrait-Multimethod-Analyse zu demonstrieren. Winne & Marx legten 181 Vpn drei Fragebogen zum Selbstbild vor. Jeder der drei Fragebogen (A = Sears' Self-Concept Inventory, B = eigener Fragebogen mit Selbsteinstufungen auf Rating-Skalen, C = eigener Fragebogen mit Vergleichen zu anderen Studenten) enthält eine Skala zu denselben drei Aspekten des Selbstbildes: "Academic", "Physical" und "Social Self-Concept", bezeichnet als Traits 1,2,3. Tabelle 4.3 gibt die Korrelationen zwischen den 3 mal 3 Fragebogenskalen an.

Tabelle 4.3: Korrelationen zwischen 3 mal 3 Fragebogenskalen zum Selbstbild. Daten von Winne & Marx (1981) zitiert nach Schwarzer (1983).

	Al	A2	A3	B1	B2	B3	C1	C2	C3
A1	1.00								
A2	.31	1							
A3	.48	.54	1						
B1	.49	-.03	-.03	1.					
B2	.22	.77	.33	.14	1.				
B3	.11	.35	.37	.06	.54	1.			
C1	.61	-.01	.10	.60	-.02	-.05	1.		
C2	.23	.13	.42	-.02	.70	.39	.14	1.	
C3	.22	.44	.55	-.07	.40	.48	.08	.56	1.

A = Sears' Self-Concept Inventory, B = Selbsteinstufung auf Rating-Skalen, C = Selbsteinstufung im Vergleich zu anderen Studenten.
1 = Academic 2 = Physical 3 = Social Self-Concept

Schwarzer paßte an diese Korrelationsmatrix ein Modell mit 3 Trait-Faktoren und 3 Methoden-Faktoren an. Dabei ließ er Korrelationen zwischen den Trait-Faktoren untereinander und zwischen den Methoden-Faktoren untereinander, nicht aber zwischen Trait- und Methoden-Faktoren zu. Das Ergebnis ist in Tabelle 4.4 angegeben.

Dieses Modell erwies sich als den Daten gut angepaßt (der Chi-Quadrat-Test für die Signifikanz der Abweichungen vom Modell ergab einen Chi-Quadrat-Wert von 10.74 bei 12 Freiheitsgraden und war nicht signifikant).

Unter inhaltlichen Gesichtspunkten fallen vor allem die negativen Korrelationen des Trait-Faktors "Academic Self-Concept" zu den Trait-Faktoren "Physical" und "Social Self-Concept" auf. Sie legen eine Interpretation im Sinne kornpensatorischer Bestrebungen nahe und sind umso bemerkenswerter, als bei jedem der drei Fragebögen die Korrelationen zwischen den drei Aspekten des Selbstbildes positiv sind. Schwarzer (1983, S.226) bemerkt dazu: "Only the structural equation approach reveals that true interrelationships between underlying sources of covariation".

Auf der Suche nach einer Alternativerklärung betrachten wir nochmals die Korrelationen in Tabelle 4.3. Es fällt auf, daß sich die Korrelationsmuster der drei Traits in den verschiedenen Fragebogenarten nahezu mustergültig wiederholen, daß dabei nur Skala Al aus dem Rahmen fällt, indem sie generell zu hoch korreliert. Deshalb wurde eine Modell-Variante konzipiert, in der Al keine reine Academic Self-Concept-Skala ist, sondern die anderen beiden Aspekte mitenthält. Es wurden deshalb Nebenladungen der Variablen Al auch auf den Trait-Faktoren 2 und 3 zugelassen. Als Ausgleich für die zusätzlichen 2 Parameter wurden die zuvor negativen Korrelationen des Trait-Faktors "Academic Self-Concept" zu den beiden anderen Trait-Faktoren auf Null fixiert. Das Ergebnis ist in Tabelle 4.5 angegeben.

Tabelle 4.4: Schwarzers (1983) Multitrait-Multimethod Lösung für die Daten aus Tabelle 4.3

Matrix der Faktorladungen

		Trait-Faktoren			Methoden-Faktoren		
		1	2	3	A	B	C
Frage-	A1	.40	0	0	.82	0	0
bogen-	A2	0	.74	0	.61	0	0
Skalen	A3	0	0	.38	.72	0	0
	B1	.63	0	0	0	.48	0
	B2	0	.70	0	0	.68	0
	B3	0	0	.48	0	.58	0
	C1	.72	0	0	0	0	.58
	C2	0	.63	0	0	0	.64
	C3	0	0	.53	0	0	.64

Korrelationen der Faktoren

	1	2	3	A	B	C
1	1					
2	-.47	1				
3	-.75	.40	1			
A	0	0	0	1		
B	0	0	0	.58	1	
C	0	0	0	.70	.57	1

Tabelle 4.5: Ergebnis einer Reanalyse der Korrelationen aus Tabelle 4.3

Matrix der Faktorladungen

		Trait-Faktoren			Methoden-Faktoren		
		1	2	3	A	B	C
Frage-	A1	.71	.18	.21	41	0	0
bogen-	A2	0	.92	0	27	0	0
Skalen	A3	0	0	.63	71	0	0
	B1	.73	0	0	0	.37	0
	B2	0	.84	0	0	.50	0
	B3	0	0	.61	0	.44	0
	C1	83	0	0	***0***	0	.41
	C2	0	77	0	0	0	.55
	C3	0	0	74	0	0	.38

Korrelationen der Faktoren

	1	2	3	A	B	C
1	1					
2	0	1				
3	0	.63	1			
A	0	0	0	1		
B	0	0	0	.03	1	
C	0	0	0	.24	.31	1

Auch dieses Modell ist gut an die Daten angepaßt. Es weist bei ebenfalls 12 Freiheitsgraden sogar noch einen etwas kleineren Chi-Quadrat-Wert aus.

Gegenüber Schwarzers Lösung fallen die Ladungen in den Trait-Faktoren höher, in den Methodenfaktoren niedriger aus und führen damit zu einem insgesamt günstigeren Urteil über die Validität der Fragebogen. Für die inhaltliche Interpretation wesentlich ist der Wegfall der negativen Korrelationen bei den Trait-Faktoren, so daß kein Anlaß zur Annahme kornpensatorischer Mechanismen (etwa entsprechend dem Klischee von den dummen Schönen und häßlichen Intellektuellen) besteht.

Mit diesem Beispiel sollte deutlich gemacht werden, daß auch mit einer konfirmatorischen Faktorenanalyse die Modellgeltung nicht bewiesen werden kann, sondern Interpretationsmöglichkeiten aufgezeigt werden. Bei hoch restriktiven Modellen wird es allerdings sehr schwer sein, gleichwertige Alternativen zu finden und damit die Interpretation in Frage zu stellen. Damit ist die Beweiskraft einer konfirmatorischen Faktorenanalyse zwar auch begrenzt, aber doch wesentlich besser als die einer klassischen Faktorenanalyse, wo Alternativlösungen routinemäßig hergestellt werden können.

Zusammenfassung

Im klassischen faktorenanalytischen Modell mit mehreren gemeinsamen Faktoren gehen individuelle Unterschiede in den Testwerten auf individuelle Unterschiede in mehreren latenten Dimensionen (= Faktoren, z.B. Fähigkeiten) zurück. Der Testwert wird als gewichtete Summe der gemeinsamen Faktoren plus einem für den jeweiligen Test spezifischen Anteil gedacht. Aufgrund der Korrelationen zwischen den Tests als Ausgangsdaten sollen die gemeinsamen Faktoren und ihr relatives Gewicht für die einzelnen Tests (= die Ladungen) bestimmt werden.

Neben der mathematischen Uneindeutigkeit der Lösung (Rotationsproblem, Kommunalitäten-Schätzproblem) haben vor allem eine Reihe weiterer Kritikpunkte, die Ende der Sechzigerjahre vorgetragen wurden (Unüberprüfbarkeit des theoretischen Ansatzes, Populationsabhängigkeit der Ergebnisse, Artefakte durch Selektion und simultane Überlagerung), dazu geführt, daß der Anspruch, Ergebnisse von Faktorenanalysen könnten als funktional erklärende Theorien über das Zustandekommen der Testwerte interpretiert werden, aufgegeben wurde. Unter Zurücknahme des ursprünglichen Anspruchs, wird die Faktorenanalyse nunmehr als Daten explorierendes, Hypothesen generierendes Verfahren eingesetzt, oder als Methode zur Definition von Beschreibungsdimensionen, oder als bloßes Datenreduktionsverfahren.

Die konfirmatorische Faktorenanalyse unterscheidet sich von der klassischen dadurch, daß der Forscher bereits Hypothesen über die Zahl der Faktoren, das Ladungsmuster, die Korrelationen der Faktoren usw. haben muß. Wenn die Hypothesen restriktiv genug sind, kann ihre Vereinbarkeit mit der empirischen Korrelations- oder Kovarianzmatrix geprüft werden. Dazu wurden vier Beispiele aus dem Bereich der Testtheorie dargestellt. Wie an Beispiel 4 gezeigt, schließt aber auch ein gut angepaßtes, hoch restriktives Modell nicht aus, daß für dieselben Korrelationen plausible Alternativerklärungen gefunden werden.

Einführende Literatur:

Bortz, J. (1989). *Statistik für Sozialwissenschaftler*. (3. Aufl.). Kapitel 15: Faktorenanalyse. Berlin: Springer.

Weiterführende Literatur:

Pawlik, K. (1971). *Dimensionen* des *Verhaltens. (2.* Aufl.). Bern: Huber.
Revenstorf, D. (1980). *Faktorenanalyse.* Stuttgart: Kohlhammer.
McDonald, R.P. (1985). *Factoranalysis and related methods.* Hillsdale: Erlbaum Ass.
Bernstein, I.H. (1987). *Applied multivariate analysis.* Chapter 7: Confirmatory factor analysis (pp. 198-245). New York: Springer.

Ein inhaltliches Beispiel, bei dem die einzelnen Schritte bei der Planung einer Faktorenanalyse detailliert dargestellt sind, findet man bei:

Rost, D.H.(1987). Leseverständnis oder Leseverständnisse? *Zeitschrift für Pädagogische Psychologie,* 1, 175-196.

4.3 Einsatzmöglichkeiten und Grenzen der Clusteranalyse

1. Wozu dienen Clusteranalysen?
2. Welche Ausgangsdaten werden benötigt ?
3. Wie können mit Hilfe von Clusteranalysen Klassifikationen erstellt werden?

Vorstrukturierende Lesehilfe

Ziel von Clusteranalysen ist es, eine Klassifikation von Objekten zu erstellen, wobei Objekte, die in dieselbe Klasse eingeordnet werden, einander möglichst ähnlich, die Klassen untereinander aber möglichst unähnlich sein sollen. Solche Aufgabenstellungen kommen in verschiedensten Wissenschaftsbereichen vor (u.a. Psychologie, Biologie, aber auch z.B. Bibliothekswissenschaften), woraus sich eine Vielfalt sich überschneidender Ansätze und Verfahren entwickelt hat, die sich ihrerseits nicht leicht in Klassen ordnen läßt. Im folgenden wird weder ein vollständiger Überblick angestrebt, noch werden einzelne Verfahren im Detail dargestellt. Es sollen lediglich die Grundgedanken skizziert und Hinweise auf mögliche Anwendungen gegeben werden. Dabei kommen im Zusammenhang mit psychologisch diagnostischen Fragestellungen vor allem zwei Anwendungsbereiche in Betracht:

(a) die Clusteranalyse von Personen als "Objekten", mit dem Ziel, möglichst homogene Personengruppen zu bilden (z.B. um in der Folge zu untersuchen, ob sich diese Gruppen in ihrer Reaktion auf eine Behandlung unterscheiden) und

(b) die Clusteranalyse von Testaufgaben, um homogene Aufgabengruppen zu finden, aus denen sich Testskalen entwickeln lassen.

Ausgangspunkt der Clusteranalyse sind Ähnlichkeitsmaße. Will man z.B. Personen zu Clustern zusammenfassen, so hat man zunächst die Ähnlichkeit (oder Unähnlichkeit, Distanz) von jeder Person zu jeder anderen festzustellen. Dazu kommen direkte Ähnlichkeitsbeurteilungen in Betracht (so könnte z.B. der Lehrer die Ähnlichkeit jedes Schülers zu jedem anderen auf einer Punkteskala beurteilen) oder auch Ähnlichkeitswerte, die aufgrund von Merkmalsausprägungen errechnet werden. Sollen z.B. Schüler nach Ähnlichkeit ihrer Interessen gruppiert werden, so könnte das Ausgangsmaterial ein standardisierter Interessentest mit zehn Unterskalen für zehn verschiedene Interessenrichtungen sein. Die Unähnlichkeit zwischen zwei Schülern könnte dann z.B. als quadrierte *euklidische Distanz* bestimmt werden: Auf jeder Interessenskala wird die Differenz bestimmt, quadriert und über alle Skalen aufaddiert. Euklidische Distanzen haben zwar den Vorteil einer anschaulichen geometrischen Bedeutung, doch kommen andere Distanzmaße oft ebensogut in Betracht: Wenn man z.B. die Differenzen nicht quadriert, sondern einfach dem Betrag nach aufaddiert, so entspricht das dem sog. *City-block-Abstand.* Bezüglich weiterer Distanzmaße und der Definition von Ähnlichkeitsmaßen aufgrund von nur rangskalierten oder nominalskalierten Merkmalen sei auf die am Ende des Kapitels genannten Lehrbücher verwiesen.

Ist die Ähnlichkeit (bzw. Distanz) von jeder Person zu jeder anderen, allgemeiner von jedem Objekt zu jedem anderen, bestimmt, so soll als Nächstes die bestmögliche

Gruppenaufteilung gefunden werden. Dazu stehen verschiedene Verfahren zur Verfügung: Bei hierarchisch agglutinierenden Clusterverfahren werden, ausgehend von der maximalen Anzahl von Clustern (d.h. jede Person wird als Cluster der Größe Eins aufgefaßt), Cluster schrittweise zusammengefaßt. Es werden zunächst die beiden Personen, die zueinander den geringsten Abstand haben, zu einem Cluster der Größe Zwei zusammengefaßt, dann wird erneut gesucht, welche beiden Cluster zueinander den geringsten Abstand haben und diese beiden zusammengefaßt, bis schließlich nur noch zwei Cluster vorhanden sind, die im letzten Schritt in eines zusammengefaßt werden. Bei jedem Schritt dieser Prozedur muß das Distanzkriterium (zulässige Distanz zwischen zwei Clustern, die zusammengefaßt werden sollen) ein Stück gelokkert werden, und man bricht die Prozedur ab (bzw. entscheidet sich im nachhinein für diese Aufteilung), wenn eine weitere Zusammenfassung einen besonders großen Schritt in der Lockerung des Distanzkriteriums erfordern würde.

Dieser Grundgedanke hierarchisch agglutinierender Clusterverfahren ist in einer Vielfalt von Algorithmen realisiert, die sich u.a. darin unterscheiden, wie der Abstand zwischen Clustern gemessen wird. Zunächst ist ja nur der Abstand zwischen Einzelobjekten, z.B. Einzelpersonen, definiert. Der Abstand zwischen zwei Clustern kann z.B. definiert werden

(a) als der kleinste Abstand zwischen einer Person aus Cluster A und einer Person aus Cluster B (Single linkage), oder

(b) als der größte Abstand zwischen zwei Personen aus A und B (complete linkage), oder auch

(c) als der mittlere Abstand (arithmetisches Mittel oder Median) aller Abstände zwischen Personen aus A und B.

Diesen Definitionen ist gemeinsam, daß sie alle aus den Abständen zwischen den Einzelobjekten (hier: Personen) errechnet werden und nicht auf die Merkmalsausprägungen zurückgreifen. Sie sind deshalb auch dann anwendbar, wenn die Ausgangsdaten beispielsweise globale Ähnlichkeitsurteile über Personen sind oder, wie bei der Clusteranalyse von Items, Korrelationen als Ähnlichkeitsmaße verwendet werden.

Wenn die Ähnlichkeit zwischen den Einzelobjekten aus Merkmalsausprägungen errechnet wurde, z.B. aufgrund von Testwerten als euklidische Distanz, so liegt es nahe, ein Cluster durch die durchschnittliche Merkmalsausprägung der darin enthaltenen Objekte zu kennzeichnen (das *Zentroid*) und die Abstände zwischen Clustern als Abstand zwischen den Zentroiden zu bestimmen. Die Heterogenität innerhalb eines Clusters kann auch als Merkmalsvarianz (Summe der Varianzen der einzelnen Merkmale oder multivariate Varianzmaße) definiert werden, und als Kriterium einer guten Clusterlösung kann definiert werden, daß die Varianz innerhalb der Cluster im Vergleich zur Varianz zwischen den Clustern (errechnet aus den Abständen zwischen den Clustermittelwerten) möglichst gering sein soll.

Neben den hierarchisch agglutinierenden Algorithmen kommen auch nicht hierarchische Verfahren zum Einsatz. Dabei wird die Clusterzahl als bekannt vorausgesetzt und, ausgehend von einer groben Näherungslösung, jedes Element probeweise in ein anderes Cluster verschoben, um zu sehen, ob sich eine Verbesserung der Clusterlösung im Sinne eines der oben genannten Kriterien ergibt. Durch das Verschieben einzelner Elemente ergibt sich eine Neudefinition der Cluster, die Abstände werden neu berechnet und es wird mit dem Verschieben fortgefahren, bis sich keine weitere Verbesserung mehr ergibt. Vielfach werden auch beide Typen von Algorithmen miteinander verbunden, indem zunächst mit hierarchisch agglutinierenden Verfahren eine

Ausgangslösung gesucht und die Clusterzahl festgesetzt wird und danach mit nicht hierarchischen Verfahren noch nach Verbesserungsmöglichkeiten gesucht wird.

Für Forschungsvorhaben im Bereich der pädagogisch-psychologischen Diagnostik kommen, wie schon eingangs erwähnt, vor allem zwei Einsatzbereiche für Clusteranalysen in Betracht: die Clusteranalyse von Personen mit dem Ziel, homogene Personengruppen zu definieren, z.B. um sie als Kriteriumsgruppen bei einer Testvalidierung zu verwenden. Mittels Clusteranalyse gefundene Kategorisierungen könnten aber auch als unabhängige Variable in Versuchsplänen herausgezogen werden, bei denen es darum geht, Behandlungseffekte weiter zu analysieren. Als zweiter Einsatzbereich ist die Clusteranalyse von Items zu sehen, mit dem Ziel aus einer großen Menge von Aufgaben Untergruppen zu bilden, aus denen sich möglichst unabhängige Skalen bilden lassen.

Wenn die Clusteranalyse, verglichen mit anderen multivariaten Verfahren, seltener zum Einsatz kommt, so dürften dafür folgende Gründe verantwortlich sein:

(a) Die Durchführung einer Clusteranalyse erfordert sowohl bei der Auswahl des Ähnlichkeitsmaßes als auch bei der Auswahl der Algorithmen und der Festlegung der Clusterzahl eine Reihe von Entscheidungen, die inhaltlich oft schwer zu begründen sind.

(b) Bei der Clusteranalyse von Personen ist immer zu bedenken, daß die untersuchten Personen nur eine Stichprobe aus der Population sind, über die Aussagen gemacht werden soll. Über den Einfluß von Stichprobenfehlern auf Clusterlösungen ist aber bislang nur wenig bekannt.

(c) Die Clusteranalyse von Items dient im wesentlichen denselben Zielen wie die Faktoranalyse. Neben der starken Tradition der Faktoranalyse ergab sich von daher kein besonderer Bedarf nach Clusterverfahren als Alternative - zumal in beiden Fällen die Korrelationen die Ausgangsbasis bilden und somit alle Einwände gegen die Faktorenanalyse, die die Populationsabhängigkeit und mögliche Artefakte bei der Berechnung von Korrelationen betreffen, für die Clusteranalyse genauso zutreffen.

Zusammenfassung

Clusteranalysen haben zum Ziel, Objekte so zu gruppieren, daß Objekte, die in dieselbe Gruppe (= Cluster) fallen, möglichst ähnlich, die Gruppen untereinander möglichst unähnlich sind. Es steht eine Vielzahl von Verfahren zur Verfügung, die sich danach unterscheiden, wie Ähnlichkeit bestimmt wird und nach welchen Algorithmen die Gruppenzusammenfassung erfolgt. In der psychologisch diagnostischen Forschung können Clusteranalysen zur Gruppierung von Personen oder auch im Rahmen der Testkonstruktion zur Gruppierung von Aufgaben zum Einsatz kommen.

Einführende Literatur:

Bortz, J. (1989). *Statistikfür Sozialwissenschaftler* (3. Aufl.). Kapitel 16: Clusteranalyse. Heidelberg: Springer.

Weiterführende Literatur:

Eckes, T. & Roßbach, H. (1980). Clusteranalysen. Stuttgart: Kohlhammer.
Krauth, J. (1983). Typenanalyse. In J. Bredenkamp & H. Feger (Hrsg.), *Strukturierung und Reduzierung von Daten* (S. 440-496). Göttingen: Hogrefe.
Oldenbürger, H.A. (1983). Clusteranalyse. In J. Bredenkamp & H. Feger (Hrsg.), *Strukturierung und Reduzierung von Daten* (S.390-439). Göttingen: Hogrefe.
Steinhausen, D. & Langer, K. (1977). *Clusteranalyse.* Berlin: de Gruyter.

5. Anforderungen an die klassischen Gütekriterien bei der Verwendung von Tests in der Forschung

Ist es gerechtfertigt, niedrigere Anforderungen an die Testgütekriterien zu stellen, wenn ein Test "nur" zu Forschungszwecken eingesetzt werden soll?

Vorstrukturierende Lesehilfe

Die klassischen Gütekriterien, nämlich Objektivität, Reliabilität, Validität und Normierung, werden auf ihre Relevanz für den Fall untersucht, daß es nicht um die Diagnose von Einzelindividuen, sondern um den Vergleich von Gruppenmittelwerten geht. Es werden verschiedene Fehlerkomponenten unterschieden und jeweils gefragt, ob sich die Fehler mit zunehmendem Stichprobenumfang ausgleichen und damit kontrollieren lassen, oder ob sie zu systematischen Unterschieden zwischen den Gruppen beitragen und damit das Ergebnis der Untersuchung verfälschen können.

5.1 Reliabilität, Objektivität, Validität

Wenn man von pädagogisch-psychologischer Diagnostik spricht, hat man als Anwendungsbereich wohl primär die Untersuchung und Beratung einzelner Schüler im Auge. Hier kommt es auf die Genauigkeit der Messung bei einem einzelnen Probanden an. Es gibt allerdings auch wichtige Anwendungsbereiche, bei denen es nicht um die Diagnostik bei Einzelindividuen, sondern um diagnostische Kennwerte für bestimmte Personengruppen oder Populationen geht. Das kann z.B. im Schulalltag der Fall sein, wenn eine bestimmte Klasse beurteilt werden soll, oder auch im Rahmen von Forschungsfragestellungen, die auf Gruppenvergleiche abzielen. Daraus ergibt sich die Frage, wie es um die Meßgenauigkeit eines Gruppenmittelwerts im Vergleich zu einem Einzelwert bestellt ist, und welche Anforderungen an die Testgütekriterien zu stellen sind, wenn es in der diagnostischen Fragestellung primär um Gruppenmittelwerte geht. Diese Frage soll im folgenden anhand eines Beispiels behandelt werden.

Wir nehmen an, jemand wolle untersuchen, ob Verbalisieren ("lautes Denken") die Leistung bei Problemlöseaufgaben verbessert. Die Versuchspersonen werden nach dem Zufall auf zwei Gruppen aufgeteilt, wovon die eine mit, die andere ohne Verbalisierung während des Problemlösens arbeitet. Die Leistungen werden dann, z.B. mithilfe des t-Tests, verglichen. Versuchsplan und Auswertung werfen hier keine besonderen Probleme auf. Vielmehr soll es im folgenden die Frage nach den Anforderungen gehen, die an den Problemlösetest hinsichtlich der klassischen Gütekriterien zu stellen sind.

Reliabilität: Wenn es um den Vergleich von Gruppenmittelwerten geht, kann man sich hinsichtlich der Reliabilität mit wesentlich geringeren Werten zufrieden geben als bei der individuellen Diagnostik. Sofern es sich um unsystematische Meßfehler (Zufallseinflüsse) handelt, kann nämlich mangelnde Meßgenauigkeit bei der einzelnen Messung durch eine Erhöhung des Stichprobenumfangs ausgeglichen werden. Das sieht man am besten, wenn man die Konfidenzintervalle für den wahren Wert vergleicht:

Bei einem einzelnen Probanden v, der einen beobachteten Wert X_v hat, lautet das Konfidenzintervall für den wahren Wert von τ_v :

$$\tau_v = X_v \pm 1.96\,\sigma(F) \qquad \text{für } \alpha = 0.05 \qquad \text{(vgl. Kapitel 2.2)}$$

Hat man es mit einer Gruppe von Probanden zu tun, die einen Mittelwert $\overline{X}$ erzielt haben, so lautet des Konfidenzintervall für den durchschnittlichen wahren Wert dieser(!) Probandengruppe (nicht zu verwechseln mit dem μ der Population, aus der diese Probanden gezogen sind):

$$\overline{\tau} = \overline{X} \pm 1.96\,\sigma(F)/\sqrt{n}$$

Man sieht, daß bei großem Stichprobenumfang n dieses Konfidenzintervall sehr eng wird, auch wenn die Fehlervarianz des Tests zunächst groß ist. Ein Test, der für die individuelle Diagnostik wegen mangelnder Meßgenauigkeit nicht mehr in Frage kommt, kann für einen Gruppenmittelwert immer noch eine zufriedenstellende Genauigkeit liefern.

Es ist jedoch zu beachten, daß die o.a. Formel lediglich die Frage beantwortet, in welchem Bereich der durchschnittliche wahre Wert dieser speziellen Probandengruppe zu suchen ist, nicht aber die Frage nach dem Mittelwert der Population, aus der diese Probanden als Zufallsstichprobe gezogen sind. Das Konfidenzintervall zu der letzteren Fragestellung lautet bekanntlich (die Ableitung findet sich in einführenden Statistikbüchern, z.B. Bortz, 1989, Kapitel 3)

$$\mu = \overline{X} \pm 1.96\,\sigma(X)/\sqrt{n}$$

D.h., die Ungenauigkeit, die entsteht, wenn der Gruppenmittelwert zur Schätzung des Populationsmittelwerts verwendet wird, hängt von der gesamten Testvarianz, also Fehlervarianz plus wahrer Varianz, ab.

Beispiel 5.1: Reliabilitätsanforderungen bei Aussagen über Einzelindividuen und Aussagen über Gruppenmittelwerte

Ein Test habe eine Varianz von 100 und eine Reliabilität von 0.9. Man vergleiche die Breite folgender Konfidenzintervalle ($\alpha = .05$):
(1) für den wahren Wert eines einzelnen Probanden
(2) für den Durchschnitt der wahren Werte einer bestimmten Gruppe von 100 Probanden
(3) für den Mittelwert der Grundgesamtheit, aus der die 100 Probanden zufällig gezogen sind. Man berechne weiter, wie sich die Breite dieser Konfidenzintervalle ändern würde, wenn es gelänge, bei gleichbleibender wahrer Varianz die Fehlervarianz auf ein Viertel zu reduzieren.

Lösung: (1) Man berechnet zunächst die Fehlervarianz nach Formel [2.7]

$$\sigma^2(F) = \sigma^2(X)\,(1 - Rel)$$

$$\sigma^2(F) = 100(1 - .9) = 10; \qquad \sigma(F) = 3.16$$

und erhält das Konfidenzintervall für den wahren Wert eines Probanden:

$$\tau_v = X_v \pm 1.96 \cdot 3.16 = X_v \pm 6.2$$

(2) Für den Mittelwert der wahren Werte einer bestimmten Gruppe von 100 Probanden erhält man:

$$\bar{\tau} = \bar{X} \pm 1.96 \cdot 3.16/\sqrt{100} = \bar{X} \pm 0.62$$

Man sieht, bei gleicher Reliabilität ist die Meßgenauigkeit für einen Mittelwert wesentlich höher als für einen Einzelwert.
(3) Hier geht es nicht nur um den Meßfehler bei der Messung dieser speziellen Probanden, sondern auch um den Stichprobenfehler bei der Ziehung der 100 Probanden aus der Grundgesamtheit. Die Breite des Konfidenzintervalls hängt von der gesamten beobachteten Varianz (d.i. wahrer Varianz plus Fehlervarianz) ab. Man erhält:

$$\mu = \bar{X} \pm 1.96 \cdot 10/\sqrt{100} = \bar{X} \pm 1.96$$

also ein wesentlich breiteres Intervall als bei Fragestellung (2)

Genauigkeitsgewinn bei Reduktion der Fehlervarianz: Indem man für die Fehlervarianz statt 10 nunmehr 10/4 = 2.5 einsetzt, sieht man, daß die Reduktion der Fehlervarianz auf ein Viertel die Breite des Konfidenzintervalls bei Fragestellung (1) und (2) halbiert. Dagegen wirkt sich bei Fragestellung (3) die Reduktion der Fehlervarianz nur wenig aus. Wenn 3/4 der Fehlervarianz wegfallt, so erhält man als neue beobachtete Varianz:

$$\sigma^2(F) = 100 - 10 \cdot 3/4 = 92.5; \qquad \sigma(F) = 9.6$$

und als Konfidenzintervall für den Mittelwert der Grundgesamtheit

$$\mu = \bar{X} \pm 1.96 \cdot 9.6/\sqrt{100} = \bar{X} \pm 1.88$$

also ein nahezu unverändertes Ergebnis. Dagegen würde eine Vervierfachung des Stichprobenumfangs (n = 400 statt n = 100) die Breite dieses Konfidenzintervalls auf die Hälfte reduzieren.

Dementsprechend hängt auch die Teststärke des t-Tests, bei dem es ja auch um einen Populationsschluß geht, von der gesamten Varianz innerhalb der Gruppen (also wahrer Varianz plus Fehlervarianz) ab. Um eine hohe Teststärke zu erhalten, ist es günstig, eine geringe Varianz innerhalb der Gruppen zu haben. Letzteres ist nicht nur eine Frage der Meßgenauigkeit des Tests (geringe Fehlervarianz), sondern vor allem eine Frage der Versuchsplanung: Wählt man möglichst homogene Gruppen (z.B. nur Studenten), so wird dadurch die wahre Varianz verringert. Freilich wird dabei auch die Aussagekraft der Untersuchung auf die entsprechende Teilpopulation (Studenten) eingeschränkt. Bei Verwendung geeigneter Versuchspläne (z.B. parallelisierte Grup-

pen statt unabhängiger Gruppen) oder einer Erhöhung des Stichprobenumfangs ist es allerdings möglich, hohe Teststärke für den Mittelwertsvergleich auch bei großer Varianz innerhalb der Gruppen zu erzielen. Es besteht also kein Anlaß, in der Testtheorie Zielgegensätze zwischen Tests für die individuelle Diagnostik (hohe Reliabilität) und Tests für die Diagnostik von Treatmenteffekten (geringe wahre Varianz und damit auch geringe Reliabilität innerhalb der Gruppen) zu konstruieren, wie das in der Literatur bisweilen geschah (z.B. Popham & Husek, 1973).

Objektivität: Im vorliegenden Beispiel sollte die Frage der Objektivität nicht kritisch werden, da es möglich sein sollte, Testdurchführung und -auswertung so weit festzulegen, daß eine hinreichende Objektivität gewährleistet ist. Man kann sich freilich auch vorstellen, es ginge um komplexe Problemlöseaufgaben, die von den Versuchspersonen in freiem Beantwortungsmodus zu lösen wären, und die Qualität der Lösung würde von einem Auswerter auf einer Punkteskala eingestuft. Objektivität wäre dann sicher nicht als selbstverständlich vorauszusetzen. Damit stellt sich die Frage, wie sich die einzelnen Fehlerkomponenten, die zu mangelnder Objektivität beitragen können, beim Vergleich von Gruppenmittelwerten auswirken (die im folgenden nur inhaltlich erläuterten varianzanalytischen Ausdrücke wie Haupteffekte, Wechselwirkung, Error sind in Kapitel 6.1 formal erklärt).

Mangelnde Objektivität kann einmal darin bestehen, daß zwischen den Beurteilern Mittelwertsunterschiede bestehen, indem manche Beurteiler generell strenger sind und im Durchschnitt weniger Punkte vergeben als andere (Haupteffekt Beurteiler). Wenn aber dieselben Beurteiler beide Probandengruppen beurteilen, sollten solche systematischen Unterschiede zwischen den Beurteilern beide Gruppen in gleicher Weise betreffen und damit den Mittelwertsunterschied zwischen den Probandengruppen nicht beeinflussen. Mangelnde Objektivität kann auch auf Zufallseinflüsse in der Beurteilung zurückgehen (Error im Sinn der Varianzanalyse). Solche Zufallseinflüsse würden - wie bereits im Zusammenhang mit der Reliabilität ausgeführt - mit zunehmendem Stichprobenumfang an Bedeutung verlieren. Mangelnde Objektivität kann schließlich auch durch Wechselwirkungen Berurteiler x Proband zustande kommen. Das wäre z.B. der Fall, wenn ein Beurteiler zwar nicht generell strenger beurteilt, aber bestimmte Fehlerarten anders bewertet als die anderen Beurteiler. Solche Wechselwirkungen könnten kritisch werden, wenn diese Fehlerarten in den beiden Probandengruppen unterschiedlich häufig vorkommen. Dann könnte es tatsächlich von der Person des Auswerters abhängen, welche Gruppe besser abschneidet. Solche Fälle dürften aber doch eher seltene Ausnahmen sein.

Im Sinne der Fragestellung gefährlich sind hingegen alle suggestiven Einflüsse, denen Beurteiler unterliegen können, wenn sie die Gruppenzugehörigkeit der Probanden und den Zweck der Untersuchung kennen. Deshalb sollten in solchen Fällen, wo immer möglich, Maße herangezogen werden, die bei der Auswertung praktisch keinen Ermessensspielraum zulassen.

Validität: Validitätsmängel, die auf Zufallsfehlern beruhen und vom zu messenden Merkmal unabhängig sind, können durch eine Erhöhung des Stichprobenumfangs kompensiert werden (vgl. Reliabilität und Objektivität). Wenn allerdings der Test inhaltlich an dem vorbeigeht, was er messen soll, dann ist dieser Mangel bei der Interpretation von Gruppenmittelwerten genauso gravierend wie bei der individuellen Diagnostik. Würde z.B. der Problemlösetest nicht Problemlösen erfordern, sondern

nur Schulwissen abfragen, so wäre das Experiment dem Sinn nach hinfällig. Mangelnde Validität der verwendeten Maße ist eine der möglichen Ursachen für das Mißglücken von Experimenten. Hinsichtlich der inhaltlichen Zulänglichkeit der verwendeten Maße sind bei der Diagnostik von Gruppenunterschieden sicherlich keine geringeren Anforderungen zu stellen als bei der individuellen Diagnostik.

5.2 Normierung

Während bei der individuell beratenden Diagnostik der Vergleich mit den Normdaten meist eine wichtige Rolle bei der Interpretation der Testbefunde spielt, lassen sich in der Forschung viele Fragestellungen auch ohne Bezugnahme auf die Normdaten beantworten. Um z.B. festzustellen, welche von zwei Gruppen im Durchschnitt mehr Treffer erzielt hat, sind offensichtlich keine Normdaten erforderlich. Sofern für den Test entsprechende Normdaten zur Verfügung stehen, kann man die beiden Mittelwerte auf der entsprechenden Skala (z.B. IQ-Einheiten) ausdrucken, um so die Größe des Mittelwertsunterschieds anschaulicher werden zu lassen.

Es lassen sich aber auch Forschungsfragestellungen denken, bei denen Normdaten eingesetzt werden: Wenn in einer altersmäßig gemischt zusammengesetzten Stichprobe die Intelligenz mit der sozialen Schichtzugehörigkeit korreliert werden soll, so wird man dazu nicht die Rohwerte (Anzahl der richtig gelösten Aufgaben) nehmen, in denen sich Intelligenz- und Alterseffekte mischen, sondern die IQ verwenden, die das unterschiedliche Alter berücksichtigen. Dabei werden sich etwaige Ungenauigkeiten in den Testnormen auf Angehörige aller sozialen Schichten etwa gleich auswirken. Daß es zu nennenswerten systematischen Verzerrungen kommt, ist zwar möglich (wenn z.B. die Normen für Zehnjährige nach oben verzerrt wären, und z.B. gerade bei den Mittelschichtkindern besonders viele Zehnjährige erfaßt worden wären), aber nicht sehr wahrscheinlich und bei sorgfältiger Versuchsplanung (gleiche Alterszusammensetzung bei allen sozialen Schichten) weitgehend vermeidbar. Der Qualität der Testnormen kommt also auch bei diesem Beispiel bei weitem keine so zentrale Bedeutung zu, wie das bei der individuellen Diagnostik der Fall zu sein pflegt.

Zusammenfassung

Wenn es nicht um die Diagnostik von Einzelindividuen geht, sondern um Aussagen über Gruppenmittelwerte und Vergleiche zwischen Gruppenmittelwerten, so können unsystematische, d.h. von Meßwert zu Meßwert unabhängige zufällige Fehler durch Vergrößerung des Stichprobenumfangs ausgeglichen werden. Im Unterschied dazu werden Validitätsmängel, die auf mangelnder inhaltlicher Zulänglichkeit beruhen, und solche Fehler, die sich auf die einzelnen Gruppen unterschiedlich auswirken, mit zunehmendem Stichprobenumfang nicht ausgeglichen, sondern gefährden die inhaltliche Interpretation und damit den Sinn der Untersuchung.

6. Weiterentwicklungen im Rahmen des klassischen Ansatzes

6.1 Die Theorie der Generalisierbarkeit

1. Was versteht man unter dem globalen wahren Wert, dem globalen Meßfehler und der globalen Reliabilität?
2. Wie kann die globale Reliabilität geschätzt werden?
3. Welche Anwendungsbereiche kommen für die Theorie der Generalisierbarkeit primär in Betracht?

Vorstrukturierende Lesehilfe

Zunächst wird der Ansatz der klassischen Testtheorie begrifflich erweitert, indem der Begriff der Testfamilie eingeführt wird, und darauf aufbauend werden die Begriffe des globalen wahren Werts, des globalen Meßfehlers und der globalen Reliabilität definiert. Danach wird ein varianzanalytischer Versuchsplan skizziert, der die Schätzung der globalen Reliabilität erlaubt. Abschließend wird auf Anwendungsmöglichkeiten hingewiesen.

6.1.1 Grundgedanken der Theorie der Generalisierbarkeit

Die Theorie der Generalisierbarkeit ist eine Verallgemeinerung der klassischen Testtheorie. Im Mittelpunkt steht der Begriff der Generalisierbarkeit, der als erweiterte Fassung des klassischen Reliabilitätsbegriffs aufzufassen ist. Die Grundgedanken wurden von Tryon (1957), Cronbach, Rajaratnam & Gleser (1963), Lord (1964), Rajaratnam, Cronbach & Gleser (1965) entwickelt; zusammenfassende Darstellungen findet man bei Lord & Novick (1968, Kapitel 8 und 9), oder Fischer (1974, Kapitel 6; 1986).

Das Interesse an einer verallgemeinerten Theorie, die mit schwächeren Annahmen auskommt als die klassische Testtheorie, läßt sich von verschiedenen Seiten her begründen: Um die Reliabilität gemäß der klassischen Testtheorie bestimmen zu können, benötigt man parallele Messungen. Da man in der praktischen Anwendung im-

mer davon ausgehen muß, daß die formale Definition der Parallelität (perfekte Übereinstimmung der wahren Werte, Gleichheit der Meßfehlerverteilungen) nicht genau erfüllt ist, stellt sich die Frage, was eigentlich geschätzt wird, wenn nicht genau parallele Tests wie parallele behandelt werden. Der Begriff der Generalisierbarkeit hat aber auch enge Beziehungen zum Begriff der Validität, wenn diese primär als inhaltliche Validität oder als Übereinstimmungsvalidität aufgefaßt wird, wie das z.B. bei lehrzielorientierten Tests der Fall ist. In der Frage nach der Generalisierbarkeit der Testergebnisse auf andere Tests mit ähnlichem Gültigkeitsanspruch geht die Frage nach der Reliabilität in die Frage nach der Validität über.

In der Theorie der Generalisierbarkeit tritt an die Stelle des Begriffs der Paralleltests der Begriff der ***"Testfamilie"*** oder der ***"nominell parallelen"*** Tests. Was man als Testfamilie definiert, ist - formal gesehen - beliebig. In der Absicht, etwas inhaltlich Sinnvolles zu definieren, wird man inhaltlich ähnliche Tests (z.B. Schulleistungstests zum selben Unterrichtsstoff) in vergleichbaren Skaleneinheiten ausgedrückt als "nominell parallel" zusammenfassen. Der ***globale wahre Wert*** eines Probanden v, bezeichnet mit ζ_v (ζ = griechisch: zeta), ist als der Durchschnitt (Erwartungswert) der wahren Werte definiert, die der Proband in den Tests der Testfamilie hat:

$$[6.1] \qquad \zeta_v = \underset{i}{E} (\tau_{vi})$$

i = Index für die Tests; die Tests werden zufällig gezogen

Im Unterschied zum globalen wahren Wert in der Testfamilie (z.B. verschiedenen Formen eines Schulleistungstests) wird dann der wahre Wert in einem bestimmten Test (z.B. Testform A) als ***"spezifischer wahrer Wert"*** bezeichnet.

Hat man einen Probanden v mit einem Test i getestet, so liegt ein beobachteter Wert X_{vi} vor. Die Abweichung dieses beobachteten Werts vom globalen wahren Wert des Probanden heißt ***"globaler Meßfehler"*** (E_{vi}).

Globaler Meßfehler: $E_{vi} = X_{vi} - C_v$

Die Abweichung vom spezifischen wahren Wert in diesem Test heißt ***"spezifischer Meßfehler"*** (F_{vi}).

Spezifischer Meßfehler: $F_{vi} = X_{vi} - \tau_{vi}$

Der Anteil der Varianz der globalen wahren Werte an der beobachteten Varianz in der Testfamilie (= beobachtete Varianz bei Zufallsziehung von Probanden und Tests) wird als ***"globale Reliabilität"*** der Testfamilie bezeichnet. Im Unterschied dazu heißt dann der Anteil der Varianz der spezifischen wahren Werte eines Tests an der beobachteten Varianz dieses Tests ***"spezifische Reliabilität"*** des Tests. Um die Varianz der globalen wahren Werte und der globalen Meßfehler zu bestimmen, legt man k Tests aus der Testfamilie einer Stichprobe von Personen vor. (Es kann hier offen bleiben, ob es sich dabei um eine Zufallsstichprobe aus der Testfamilie handelt, oder die Testfamilie nur aus diesen k Tests besteht). Ein entsprechender Versuchsplan ist in Tabelle 6.1 dargestellt.

Tabelle 6.1: Zweifaktorieller varianzanalytischer Versuchsplan zur Schätzung der globalen Reliabilität. Zeilenfaktor = Personen = A; Spaltenfaktor = Tests = B

		Faktor B = Tests 1 2 . . . . i . . . . . k	Personen mittelwerte
	1	X_{11} . . .	$X_{1.}$
Faktor A Personen	2	X_{21} . . .	$X_{2.}$
	v	. . . . . X_{vj}. . .	$X_{v.}$
	N	X_{N1} . . . . . X_{NK}	$X_{N.}$
Testmittelwerte		$X_{.1}$ $X_{.i}$ $X_{.k}$	$X_{..}$

Faßt man diesen Datenerhebungsplan als zweifaktoriellen varianzanalytischen Versuchsplan (Zeilenfaktor = Personen, Spaltenfaktor = Tests) mit einem Meßwert pro Zelle auf, so kann man den beobachteten Wert in varianzanalytische Komponenten zerlegen. An dieser Zerlegung läßt sich der Unterschied zwischen globalem und spezifischem Meßfehler inhaltlich deutlicher machen. In einem zweifaktoriellen Versuchsplan lautet die Zerlegung eines Meßwerts:

$$X_{vi} = \mu + \alpha_v + \beta_i + \alpha\beta_{vi} + res_{vi}$$

mit μ = Erwartungswert über alle Personen und Tests

α_v = Haupteffekt der Person. Haupteffekte der Personen drücken individuelle Unterschiede im globalen, d.h. über alle Tests gemittelten Leistungsniveau aus. Personen-Haupteffekte sind Unterschiede in den globalen wahren Werten.

$ß_i$ = Haupteffekt des Tests. Damit werden Schwierigkeitsunterschiede zwischen den Tests ausgedrückt.

$\alpha\beta_{vi}$ = Wechselwirkungseffekt. Abweichung des Erwartungswertes einer Zelle (= Erwartungswert einer Person v in einem Test i, also ihr spezifischer wahrer Wert τ_{vi}) von dem, was sich aufgrund der Haupteffekte (= Schwierigkeit des Tests, globales Leistungsniveau der Person) allein ergeben würde. Ein positiver Wechselwirkungsbetrag würde z.B. entstehen, wenn eine sonst durchschnittliche Person gerade den Test bekäme, auf dessen Aufgaben sie sich besonders gut vorbereitet hat. (Das Gegenteil kann auch vorkommen).

res_{vi} = Von Messung zu Messung unabhängiger Zufallseinfluß

Zerlegt man den Meßwert eines Probanden in den globalen wahren Wert und den globalen Meßfehler, so entspricht das folgender Zusammenfassung:

$$X_{vi} = (\mu + \alpha_v) + (\beta_i + \alpha\beta_{vi} + res_{vi}) = \zeta_v + E_{vi}$$

d.h. Schwierigkeitsunterschiede zwischen den Tests und Wechselwirkungseffekte werden dem globalen Fehler zugerechnet. Die Zerlegung in den spezifischen wahren Wert und spezifischen Meßfehler entspricht dagegen folgender Zusammenfassung:

$$X_{vi} = (\mu + \alpha_v + \beta_i + \alpha\beta_{vi}) + (res_{vi}) = \tau_{vi} + F_{vi}$$

d.h. Schwierigkeitsunterschiede zwischen den Tests und Wechselwirkungseffekte werden zum spezifischen wahren Wert gerechnet. Nur die unabhängigen Zufallseinflüsse zählen zum spezifischen Fehler. Die rechnerische Durchführung der Varianzanalyse und die Schätzung der Varianzanteile soll hier nicht im einzelnen dargestellt werden. Die globale Fehlervarianz in der Testfamilie läßt sich relativ einfach als Varianz der Testwerte, die von derselben Person stammen, schätzen. Die Schätzung weiterer Komponenten (Varianz der globalen wahren Werte, Aufspaltung der globalen Fehlervarianz in Anteile zu Lasten von Schwierigkeitsunterschieden, Wechselwirkungen, spezifischen Meßfehlern) richtet sich danach, ob die vorliegenden Tests die gesamte Testfamilie ausmachen oder ob sie als Zufallsstichprobe aus der Testfamilie aufgefaßt werden (Näheres dazu siehe Lord & Novick, 1968, Kapitel 7-9; Fischer, 1974, Kapitel 6.3). Eine Weiterführung des Ansatzes (Erweiterung des zweifaktoriellen varianzanalytischen Versuchsplans auf 3 und mehr Dimensionen) findet man bei Nußbaum (1987). Ein dreifaktorieller Versuchsplan entsteht z.B., wenn jede von N Personen jeden von k Tests in jeder von m Situationen bearbeitet hat.

6.1.2 Anwendungsmöglichkeiten

(a) Übereinstimmung zwischen Tests mit ähnlichem Validitätsanspruch: Tests mit ähnlichem Validitätsanpruch und gleichen Skaleneinheiten (vergleichbare Rohwertskalen oder gleiche Standardisierung) können zu einer Testfamilie zusammengefaßt werden. Die Frage nach der Übereinstimmung innerhalb einer solchen Testfamilie ist von unmittelbarem praktischem Interesse. Die globale Reliabilität kann zugleich als Maß der konvergenten Validität betrachtet werden. Die Angabe der globalen Fehlervarianz beantwortet die Frage, welche Varianz im Durchschnitt zu erwarten ist, wenn ein Proband mit den verschiedenen Tests getestet wird. Die Datenerhebung für den in Tabelle 6.1 dargestellten Versuchsplan ist allerdings recht aufwendig, da jeder Proband alle Tests bearbeiten muß. Hat man solche Daten zur Verfügung, so wird man sich auch nicht mit der globalen Charakterisierung der Testfamilie begnügen, sondern darüber hinaus die einzelnen Tests näher betrachten: Wenn z.B. Haupteffekte der Tests signifikant waren, wird man weiter fragen, welche Tests leichter oder schwerer waren. Man wird sich für die Korrelationen zwischen den Tests interessieren, um sie inhaltlich zu interpretieren. Solche Einzelergebnisse sind hier mindestens so belangvoll wie die globale Beschreibung der Testfamilie.

(b) Konstruktion von nominell parallelen Tests durch Item-Sampling: Nominell parallele Tests können auch dadurch definiert werden, daß aus einem Pool von Aufgaben jeweils k Aufgaben zufällig gezogen werden. Alle möglichen Ziehungsergebnisse, d.h. alle möglichen Tests aus k Aufgaben bilden die Testfamilie. Der globale wahre Wert des Probanden ist dann die Trefferzahl, die er im Durchschnitt über alle Testziehungen zu erwarten hat. Bei einer einzelnen Testziehung mag der Proband Glück oder Pech haben, indem er leichte oder schwierige Items zieht (=Haupteffekt Tests) oder auch Items, die gerade ihm leicht bzw. gerade ihm schwer fallen (=Wech-

selwirkung Test-Proband). Diese Effekte sowie Zufallseinflüsse bei der Bearbeitung der gezogenen Items tragen zum globalen Fehler bei. Die Theorie des Item-Samplings wurde zwar früh entwickelt (siehe Lord & Novick, 1968, Kapitel 11), aber selten praktisch zum Einsatz gebracht. Der Grund dafür dürfte darin liegen, daß entsprechende Itempools nicht zur Verfügung stehen - wenngleich verschiedene Ansätze dazu vorhanden sind (siehe Kapitel 6.2 und 7.4). Ein relativ frühes Anwendungsbeispiel findet man bei Hively, Patterson & Page (1968) die mit Hilfe der Theorie der Generalisierbarkeit die Übereinstimmung von verschiedenen Mathematiktests, die nach demselben Schema konstruiert waren (siehe Kapitel 6.2), bestimmten.

(c) Anwendungsmöglichkeiten bei der Bestimmung der Auswertungsobjektivität: Zur Feststellung der Auswertungsobjektivität eines Tests werden die Testprotokolle einer repräsentativen Stichprobe von Probanden von mehreren unabhängigen Auswertern beurteilt, sodaß der in Tabelle 6.1 dargestellte Versuchsplan realisiert ist. Formal gesehen tritt nun an die Stelle der Testfamilie die Population möglicher Auswerter und die vorhandenen Auswerter werden als Zufallsstichprobe daraus betrachtet. Als Maß der Auswertungsobjektivität wird die globale Reliabilität der Urteile berechnet. Dabei werden Mittelwertsunterschiede zwischen den Auswertern (manche Beurteiler mögen bei der Vergabe der Punkte großzügiger sein, andere strenger), Wechselwirkungen zwischen Protokoll und Beurteiler (manchen Beurteilern mögen bestimmte Arten von Antworten besonders gefallen oder mißfallen) und reine Zufallseinflüsse zu den Fehlern gerechnet und gehen in die Fehlervarianz ein. Daß all diese Komponenten zum Fehler gerechnet werden, unterscheidet die varianzanalytische Berechnung der Auswertungsobjektivität von anderen Methoden: Berechnet man z.B. einfach die durchschnittliche Korrelation zwischen den Beurteilern, so bleiben Varianzunterschiede außer Betracht, da ja Korrelationen darauf nicht reagieren. Wenn es nicht nur darauf ankommt, daß die Beurteiler die Probanden in dieselbe Rangreihe bringen, sondern auf die numerische Übereinstimmung (daß sie dieselbe Leistung mit demselben Punktwert oder derselben Note belegen), ist die varianzanalytische Bestimmung der Auswertungsobjektivität vorzuziehen. Ein frühes Anwendungsbeispiel findet man bei Michel & Mai (1969), die mit Hilfe von Varianzkomponenten-Zerlegungen die Auswertungsobjektivität verschiedener Untertests des HAWIE (Hamburg-Wechsler-Intelligenztest für Erwachsene nach Hardesty & Lauber, 1956) bestimmten.

Zusammenfassung

Eine Testfamilie ist eine endliche oder unendliche Menge von nominell parallelen Tests. Der globale wahre Wert eines Probanden ist sein durchschnittlicher wahrer Wert gemittelt über die Testfamilie; der globale Meßfehler die Abweichung eines beobachteten Testwerts vom globalen wahren Wert. Die globale Reliabilität ist der Anteil der Varianz der globalen wahren Werte an der beobachteten Testvarianz. Die einzelnen Varianzkomponenten können geschätzt werden, wenn einer Stichprobe von Personen k Tests (k = alle oder k zufällig ausgewählte) vorgelegt werden. Die Varianzzerlegung folgt dem allgemeinen Schema der Auswertung varianzanalytischer Versuchspläne. Die Theorie der Generalisierbarkeit kann verwendet werden, um die Übereinstimmung von als parallel konzipierten Testformen auszudrücken. Sie kann - in etwas anderem Kontext - auch verwendet werden, um Beurteilerübereinstimmung zu schätzen.

Fragen der Beurteilerübereinstimmung treten nicht nur im Zusammenhang mit der Auswertung von Tests auf. Auch bei der Auswertung von Verhaltens-Protokollen, Interview-Daten usw. stellen sich ähnliche Fragen. Sofern quantifizierbare Daten vorliegen, kommt auch hier eine varianzanalytsiche Berechnung der Beurteilerübereinstimmung in Betracht.

Einführende Literatur:

Fischer, G.H. (1974). ***Einführung*** in die ***Theorie psychologischer Tests.*** Kapitel 6: Theorie der Verallgemeinerung von Testergebnissen und die statistische Schätzung von Reliabilitätskoeffizienten. Bern: Huber.

Weiterführende Literatur:

Lord, EM. & Novick, M.R. (1968). ***Statistical theories of mental test scores.*** Kapitel 8: Some test theory for imperfectly parallel measurements; Kapitel 9: Types of reliability coefficients and their estimation. Reading, Mass.: Addison-Wesley.

Nußbaum, A. (1987). Das Modell der Generalisierbarkeitstheorie. In: Klauer, K.J. ***Kriteriumsorientierte Tests.*** S. 114-136. Göttingen: Hogrefe.

6.2 Kriterienorientierte versus normorientierte Messung

1. Was versteht man unter kriterienorientierter Messung?
2. Sind kriteriumsorientierte Tests nach denselben Prinzipien zu konstruieren und nach denselben Gütekriterien zu beurteilen wie normorientierte?
3. Welche besondere Rolle spielt die inhaltliche Validität und wie läßt sich inhaltliche Validität begründen?
4. Welche Annahme macht das Binomialmodell und wie läßt sich auf dieser Grundlage ein Entscheidungsmodell (Kriterium erreicht/nicht erreicht) begründen?

Vorstrukturierende Lesehilfe

Bei kriterienorientierter Messung geht es darum, die Leistung des Probanden mit einem inhaltlich definierten Anforderungskriterium, z. B. einem Lehrziel, zu vergleichen. Diese Zielsetzung wird zunächst von normorientierter Messung (Vergleich mit einer Normpopulation) abgegrenzt (6.2.1). Sodann wird die Frage behandelt, welche Bedeutung Kennwerte der klassischen Testtheorie wie Reliabilität, Validität, Standardmeßfehler bei kriterienorientierter Messung haben. Dabei wird u.a. auch der als Alternative zu den klassischen Gütekriterien vorgeschlagene Übereinstimmungskoeffizient U diskutiert und auf Mängel dieses Koeffizienten hingewiesen (6.2.2). Bei dem Anforderungskriterium, über dessen Erreichen/Nicht-Erreichen mithilfe eines kriterienorientierten Tests entschieden werden soll, handelt es sich in der Regel um ein bestimmtes Lehrziel. Damit stellt sich die Frage, ob der Test für dieses Lehrziel repräsentativ ist, d.h. ob er inhaltliche Validität besitzt. Es werden Itemkonstruktionsverfahren dargestellt und diskutiert, die inhaltliche Validität gewährleisten sollen (6.2.3).

Das Anforderungskriterium kann so definiert sein, daß der Proband Aufgaben eines bestimmten Typs mit einer bestimmten festgesetzten Wahrscheinlichkeit lösen muß. Werden Items zufällig gezogen, so kann man das Binomialmodell anwenden, um die Trefferwahrscheinlichkeit eines Probanden zu schätzen, und Entscheidungsregeln entwickeln, ab wann das Kriterium als erreicht gelten soll. Diese testtheoretischen Entwicklungen werden kurz dargestellt (6.2.3). Wenngleich ähnliche Gedanken auch schon früher geäußert wurden, so dürfte die Diskussion um kriterienorientierte versus normorientierte Tests doch ihre wesentlichen Impulse ausgehend von den Arbeiten von Ebel (1962) und Glaser (1963) erhalten und Ende der 60er/Anfang der 70er Jahre ihren Höhepunkt erreicht haben. Die Hauptergebnisse wurden von Klauer (1983; 1987) zusammenfassend dargestellt. Näheres zur Geschichte findet man auch bei Hilke (1980).

6.2.1 Die Zielsetzung kriterienorientierter Messung

Während es in der klassischen Testtheorie darum geht, individuelle Unterschiede zu erfassen, und dementsprechend das Testergebnis des Probanden gewöhnlich im Vergleich mit einer Normpopulation (z. B. den Gleichaltrigen, vgl. Kapitel 2.3) angegeben und interpretiert wird, geht es bei kriterienorientierter Messung um die Frage, ob der Proband ein bestimmtes, inhaltlich definiertes Lehrziel erreicht hat. Wieviele andere Probanden das Lehrziel ebenfalls erreicht haben, und wie sich die Leistung des Probanden von der anderer unterscheidet, steht dabei nicht zur Diskussion. Dementsprechend muß das Kriterium, anhand dessen über das Erreichen des Lehrziels entschieden wird, a priori, das heißt ohne Bezugnahme auf die Verteilung der Werte in einer Gruppe oder Population, festgesetzt werden. Das ist z. B. der Fall, wenn als Kriterium dafür, daß ein Kind das Lehrziel "Addieren im Zahlenbereich 1 bis 100" erreicht hat, festgesetzt wird, daß bei zufällig ausgewählten Additionsaufgaben nicht mehr als 5 % Fehler vorkommen dürfen. Solche Vergleiche des Leistungsstandes eines Probanden mit inhaltlich definierten Kriterien sind vor allem bei Fragen der Planung, aber auch der Erfolgskontrolle von Unterricht von Interesse.

6.2.2 Die Auseinandersetzung mit der klassischen Testtheorie

Aufgrund der speziellen Zielsetzung kriterienorientierter Messung haben zunächst einige Autoren (Fricke, 1972; 1974; Herbig, 1973) die Ansicht vertreten, die klassische Testtheorie mit ihren Prinzipien der Testkonstruktion und ihren Testgütekriterien sei für kriterienorientierte Tests ungeeignet. Die klassische Testtheorie sei entwikkelt worden, um individuelle Unterschiede zwischen Probanden zu erfassen, während es doch das Ziel eines guten Unterrichts sein müsse, zu erreichen, daß alle Probanden das Lehrziel erreichen, also diesbezüglich individuelle Unterschiede verschwinden. Wenn alle den Lehrstoff vollständig beherrschten, gäbe es keine Testvarianz mehr, und die als Korrelationen definierten Gütekriterien (Reliabilität und Validität, aber auch Trennschärfekoeffizienten für die einzelnen Items) seien nicht mehr angebbar (Fricke 1972; 1974; Ingenkamp, 1985; ähnlich Klauer, 1987). In diesem Sinn spricht z. B. Fricke (1972) von einem "Versagen" der klassischen Testtheorie bei kriterienorientierten Tests. Als Alternative zu den klassischen Gütekriterien bietet Fricke (1972) den Ü-Koeffizienten an, der die Übereinstimmung zwischen zwei oder mehreren Tests oder auch zwischen zwei oder mehreren Beurteilern ausdrücken soll. Der Ü-Koeffizient ist wie folgt definiert:

$$\ddot{U} = 1 - \frac{Var}{maxVar}$$

Var = durchschnittliche Varianz der Urteile über einen Probanden (der Testergebnisse eines Probanden)

maxVar = maximal mögliche Varianz der Urteile über einen Probanden. Sie tritt auf, wenn die Hälfte der Beurteiler den einen, die andere Hälfte den anderen Extremwert nennt (wenn der Proband bei der einen Hälfte der Tests den kleinstmöglichen, bei der anderen Hälfte den größtmöglichen Wert erzielt).

Wenn es nur um zwei Beurteiler (oder zwei Tests) geht und das Ergebnis nur in zwei Kategorien (Lehrziel erreicht/nicht erreicht) ausgedrückt wird, gibt der Ü-Koeffizient den Prozentsatz übereinstimmender Entscheidungen an. Der Ü-Koeffizient als Prozentsatz übereinstimmender Entscheidungen wurde von Fricke (1972) sowohl für Fragen der Reliabilität (Übereinstimmung zwischen zwei Tests) als auch der Validität (Übereinstimmung zwischen Test und Kriterium) empfohlen.

Als Alternative zu den Itemselektionsverfahren der klassischen Testtheorie schlägt Glaser (1973) vor, die Items auszuwählen, die den Lernfortschritt am besten sichtbar machen, indem sie von einer Personengruppe nach dem Training wesentlich häufiger gelöst werden als vorher.

Diese Kritik an der klassischen Testtheorie blieb nicht unwidersprochen (Stelzl, 1976). Zwar trifft es zu, daß Reliabilität und Validität populationsabhängig definiert sind (vgl. Kapitel 2.5), doch bietet die klassische Testtheorie mit dem Begriff des Standardmeßfehlers und den darauf aufbauenden Konfidenzintervallen (vgl. Kapitel 2.2) auch Konzepte an, die es ermöglichen, unabhängig von der Verteilung der wahren Werte die Meßgenauigkeit für den einzelnen Probanden anzugeben. Im Extremfall, wenn alle Personen denselben wahren Wert haben, reduziert sich die beobachtbare Testvarianz auf die Fehlervarianz. Nur in dem praktisch wenig realistischen Spezialfall, daß auch die Fehlervarianz Null ist, könnte die Testvarianz Null werden. Aber selbst dieser Fall führt nicht zu theoretischen Problemen: Mit der Feststellung "die Fehlervarianz ist Null" ist die Frage nach der Meßgenauigkeit ja ebenfalls beantwortet.

Der Ü-Koeffizient, der von Fricke (1972) als Alternative vorgeschlagen wurde und auch bei Ingenkamp (1985) und Klauer (1987) dargestellt ist, wurde von Stelzl (1976) kritisiert. Hier soll nur der einfachste Fall betrachtet werden, bei dem nur zwei Kategorien (Lehrziel erreicht oder nicht erreicht) unterschieden werden. Wie oben erwähnt gibt dann der Ü-Koeffizient für zwei Tests den Prozentsatz übereinstimmender Entscheidungen an. Dieses sehr einfache und anschauliche Maß erweist sich allerdings bei näherem Hinsehen als wenig geeignet, über die inhaltliche Übereinstimmung oder die Meßgenauigkeit von Tests Auskunft zu geben:

(1) Wenn zwei Tests beide so leicht sind, daß alle Probanden alle Aufgaben lösen, oder beide so schwer, daß kein Proband eine Aufgabe löst, so ergibt sich unabhängig vom Inhalt der beiden Tests, der völlig verschieden sein kann, immer eine Übereinstimmung von Ü = 1,0. Wenn z. B. in zwei Tests, die voneinander unabhängige Fähigkeiten prüfen (Diskuswerfen und Lateinvokabeln), jeweils 90 % der Schüler das Lehrziel erreichen, so ergibt sich ein Ü = 0,82 (0,9 x 0,9 = 0,81. 0,1 x 0,1 = 0,01). Als Maß dafür, inwieweit zwei Tests dasselbe Merkmal erfassen, ist der Ü-Koeffizient somit offensichtlich irreführend.

(2) Auch wenn es sich um Test und Retest handelt, so daß die Frage nach der inhaltlichen Übereinstimmung als beantwortet gelten kann, besagt ein hoher Ü-Koeffizient nicht, daß dieser Test für diese bestimmte Probandengruppe als Meßinstrument, etwa zur Erfassung eines Lernfortschrittes, geeignet sein wird. Ist der hohe Ü-Koeffizient auf extreme Testschwierigkeit (oder Leichtigkeit) zurückzuführen, so ist der Test trotz hohem Ü nicht geeignet, Lernfortschritte sichtbar zu machen.

(3) Zeigen zwei als parallel konzipierte Tests einen niedrigen Ü-Koeffizienten, so kann das an mangelnder inhaltlicher Übereinstimmung oder an mangelnder Meßgenauigkeit liegen, was schwerwiegende Mängel wären. Es kann aber - zumal bei nicht normierten Tests - auch an einer Skalenverschiebung liegen, wodurch das Kri-

terium an unterschiedlichen Stellen des Leistungskontinuums zu liegen kommt. Letzteres wäre durch eine Skalentransformation leicht zu beheben.

Während in der klassischen Testtheorie zwischen Meßgenauigkeit (Reliabilität), inhaltlicher Übereinstimmung (Vergleich der Paralleltestreliabilität mit anderen Arten der Reliabilitätsbestimmung) und Übereinstimmung der Skalierung (Normierung) unterschieden werden kann, sind diese Gesichtspunkte konfundiert, wenn man lediglich den Ü-Koeffizienten als Prozentsatz übereinstimmender Entscheidungen angibt. Der Ü-Koeffizient kann somit als Alternative zu den Koeffizienten der klassischen Testtheorie bzw. den dort definierten Fehlermaßen nicht überzeugen.

Auch die Methoden der Testkonstruktion, die die klassische Testtheorie anbietet, stehen nicht im Widerspruch zu den Anliegen kriterienorientierter Messung. Vertreter des kriterienorientierten Ansatzes fordern, für einen lehrzielorientierten Test müßten die Items so ausgewählt werden, daß ein Lernfortschritt möglichst gut sichtbar werde. Das seien solche Items, die von einer Personengruppe vor dem Unterricht mit sehr niedriger, nach dem Unterricht mit sehr hoher Wahrscheinlichkeit gelöst werden. Eine solche Itemauswahl entspricht in der Terminologie der klassischen Testtheorie einer Itemselektion nach den Itemvaliditäten unter Verwendung des dichotomen Merkmals "Unterricht absolviert: Ja/nein" als Außenkriterium, steht also nicht im Gegensatz zu den Vorgehensweisen der klassischen Testtheorie.

Inzwischen ist die erste Phase der Diskussion, in der vor allem die Unterschiede zwischen kriteriumsorientierter und normorientierter Messung betont wurden, abgeklungen. Mittlerweile dürfte sich allgemein die Auffassung durchgesetzt haben, daß es hier nicht um gegensätzliche Prinzipien der Testkonstruktion geht, die entsprechend zu zwei verschiedenen Klassen von Tests führen müßten, sondern um unterschiedliche Interpretationsweisen von Tests. So z.B. kommen Tent & Waldow (1984) nach einer grundsätzlichen Diskussion um die Funktion pädagogischer Diagnostik und einer Auseinandersetzung mit den von beiden Seiten vorgetragenen Argumenten zu dem Ergebnis, daß Gruppennorm- und Lehrzielorientierung ineinander überführbare Aspekte pädagogischer Leistungsmessung sind. Ein Test, der für einen bestimmten Lehrstoff repräsentativ ist und für den Normwerte aus geeigneten Vergleichspopulationen vorliegen, ermöglicht sowohl einen Vergleich der Leistung des Probanden mit inhaltlich definierten Standards als auch mit der Normpopulation. Dementsprechend wäre zutreffender, nicht von kriterienorientierten versus normorientierten Tests, sondern von kriterienorientierter oder normorientierter Testbefundinterpretation zu sprechen. Im Hinblick auf eine kriterienorientierte Testbefundinterpretation lauten dann die Hauptfragen an die Testtheorie: Wie bildet man aus einem Lehrstoff eine repräsentative Aufgabenmenge? Wie entscheidet man, ob das Lehrziel erreicht ist? Zu diesen Fragen wurden von verschiedener Seite Beiträge geleistet, die im folgenden kurz dargestellt werden (einen ausführlichen Überblick gibt Klauer, 1983; 1987).

6.2.3 Spezifische Probleme lehrzielorientierter Tests

6.2.3.1 Inhaltliche Validität

Bei der Konstruktion lehrzielorientierter Tests besteht der erste und entscheidende Schritt darin, Aufgaben zu konstruieren, die für den Lehrstoff repräsentativ sind, so daß für den Test inhaltliche Validität in Anspruch genommen werden kann. Dazu

wurden verschiedene Verfahren entwickelt, die Wieberg (1983) grob in "umgangssprachlich orientierte" und "formalsprachlich orientierte" einteilt. Erstere enthalten eher allgemein gehaltene Anleitungen zur Aufgabenkonstruktion, letztere versuchen für bestimmte Bereiche Aufgabenuniversa so eng zu definieren, daß die einzelnen Aufgaben auch von einem Computer generiert werden könnten.

Zu den umgangssprachlich orientierten Verfahren sind unter anderem die Lernzieloperationalisierung nach Mager (1965), die Konstruktion einer Lehrzielmatrix nach Tyler (1950) oder einer Lehrzieltaxonomie nach Bloom et al. (1971) zuzurechnen.

Nach Mager (1965) soll eine Lernzieloperationalisierung folgende Elemente enthalten:

1. Angabe über die vom Probanden am Ende des Unterrichts geforderte Tätigkeit. Dabei soll es sich um direkt beobachtbares Verhalten (Lösen quadratischer Gleichungen, Reparieren von Radios, Aufzählen von Hauptstädten) handeln. Ausdrücke, die sich auf nicht direkt beobachtbare Zustände und Prozesse wie "Wissen", "Verstehen", "Würdigen können" beziehen, sind durch Angaben über beobachtbares Verhalten zu ersetzen.
2. Angabe über die Bedingungen, unter denen das geforderte Verhalten zu zeigen ist (z. B. über erlaubte Hilfsmittel).
3. Angabe von Kriterien, unter denen das Lehrziel als erreicht gilt (z. B. geforderter Prozentsatz richtiger Lösungen).

Diese von Mager geforderte Operationalisierung von Lernzielen ist sicher geeignet, allzu vage Lernzielbeschreibungen zu konkretisieren. Für die Testkonstruktion besagt sie allerdings kaum mehr, als daß aus einem Lernziel konkrete Testaufgaben hergeleitet werden sollen, wobei die Frage, wie diese inhaltliche Umsetzung zu geschehen hat, weitgehend offen bleibt. Anleitungen zur inhaltlichen Untergliederung eines komplexen Lehrzieles wurden schon von Tyler (1950) vorgelegt. Bei der Ableitung von Teillehrzielen werden eine Inhaltsdimension (Teilung des Lehrstoffes in inhaltliche Abschnitte) und eine Handlungsdimension (z. B. Wissen, Verständnis, Anwendung) unterschieden. Wenn man die beiden Dimensionen systematisch kombiniert, erhält man die "Tyler-Matrix". Jede Zelle dieser Matrix entspricht einem Teillehrziel, für das dann Testaufgaben zu konstruieren sind (vgl. Abschnitt 10.2).

Tabelle 6.2: Schema **einer Tyler-Matrix**

Inhaltskomplexe	Wissen	Verstehen	Anwendung
A			
B			
C			

Das Grundschema der Tylermatrix wurde in den Lehrzieltaxonomien von Bloom weiterentwickelt und auf verschiedensten Wissensgebieten angewendet (Bloom et al., 1971). Auch diese Gruppe von Verfahren wird von Wieberg (1983) den umgangssprachlich orientierten zugerechnet, da sie dem Testkonstrukteur bei der inhaltlichen

Ausgestaltung viel Spielraum lassen und Tests zum selben Lehrziel je nach Testautor unterschiedlich ausfallen können.

Den Verfahren, die Wieberg (1983) als "formalsprachlich orientiert" bezeichnet, ist gemeinsam, daß Regelsysteme entwickelt werden, mit denen ein Aufgabenuniversum möglichst eindeutig festgelegt werden soll. Dazu bieten sich zunächst relativ leicht abgrenzbare Teilgebiete, z. B. aus dem Bereich der Mathematik, an. Osburn (1968) und Hively et al. (1968) definieren Aufgabenuniversa mit Hilfe von Aufgabenschemata ("Itemforms"), aus denen dann durch Einsetzen von Zahlen oder Objekten Aufgabenmengen entstehen.

Ein einfaches Beispiel für ein Aufgabenschema aus dem Bereich des Grundrechnens könnte wie folgt aussehen:

a x b = ? a und b sind natürliche Zahlen zwischen 1 und 10.

Allgemein gesprochen besteht ein solches Aufgabenschema aus der Angabe einer festen syntaktischen Struktur, die eine oder mehrere variable Elemente (hier: a und b) enthält, sowie der Angabe von Regeln, nach denen für die variablen Elemente einzusetzen ist (hier: a und b sind natürliche Zahlen von 1 bis 10), um Items (z. B. die Frage "7 x 2 = ?") zu generieren. Hively et al. (1968) konstruierten mit dieser Methode Aufgabenuniversa für verschiedene elementare Rechenarten. Hinweise auf weitere Anwendungen findet man bei Wieberg (1983) und Klauer (1987). So z.B. schlägt Klauer (1978; 1987) vor, diese Methode zur Konstruktion von Itempools für Tests aus dem klassischen Bereich der Intelligenzmessung zu verwenden, etwa zur Konstruktion verbaler Analogieaufgaben (Gras: grün = Himmel: ?). Wenn Mengen von Relationen (Teil von, größer als . . .) und zu jeder Relation Mengen von Einsetzobjekten (z. B. zur Relation "Teil von" die Wortpaare Nase-Gesicht, Henkel-Tasse usw.) definiert sind, so kann daraus ein Itempool generiert werden. Kritisch anzumerken bleibt, daß die Fähigkeit, diesen speziellen Itempool zu lösen, nicht von Interesse ist. Das Testergebnis soll vielmehr als Indikator breiter definierter Intelligenzfaktoren verwendet werden. Dementsprechend interessiert hier nicht inhaltliche Validität bezogen auf eine eng definierte Itemmenge, sondern der Wert des Tests als Indikator und Prädiktor (Übereinstimmung mit Außenkriterien, prognostische Validität usw.).

Streng regelgeleitete Methoden der Itemkonstruktion scheinen zunächst auf einen Lehrstoff wie z. B. Geschichtswissen, bei dem Textverstehen und Erfassen von komplexen inhaltlichen Zusammenhängen im Vordergrund stehen, nicht anwendbar zu sein. Trotzdem wurden auch für solche Lehrstoffe verschiedene Verfahren entwickelt, um sie mit Hilfe formaler Regeln in Aufgabenmengen umzusetzen. So z. B. empfiehlt Klauer (1987), den Lehrtext zunächst in eine Folge von Aussagen umzuschreiben, die den Sachverhalt vollständig, aber ohne Weitschweifigkeit oder Wiederholungen darstellen. Beispiel für eine solche Einzelaussage ist der Satz "Kolumbus hat Amerika im Jahr 1492 entdeckt". Diese Aussagen werden in Fragesätze umgeformt, wobei nach jedem Satzteil (wer? was? wann?) gefragt werden kann. Aus diesen Fragen kann dann ein Test zusammengestellt werden, z. B. durch Zufallsauswahl der Fragen oder nach vorheriger Untergliederung des Textes in Abschnitte und Ziehen einer bestimmten Fragenzahl aus jedem Abschnitt. Ähnliche Ansätze, die teils auf grammatischen Strukturanalysen eines vorliegenden Textes aufbauen oder aber zunächst eine Neuformulierung des Textes, z. B. eine Transformation in Propositionen erfordern, sind bei Feger (1984) zusammenfassend dargestellt und diskutiert.

Wenngleich es mit Hilfe formalsprachlich orientierter Verfahren möglich ist, gut abgegrenzte Aufgabenmengen zu erzeugen, so sollte doch nicht übersehen werden,

daß dabei eine Reihe von Entscheidungen zu treffen sind, die mehr oder weniger gut begründet sein mögen, sich aber nicht zwingend aus dem Lehrziel ableiten lassen. Ein Unterrichtsstoff wie "Ursachen und Folgen des 30jährigen Krieges" ist sicher nicht eindeutig in eine Aufgabenmenge zu zerlegen, sondern das Ergebnis wird stark vom Historiker bzw. vom Testautor abhängen. Selbst bei gegebener Aussagenmenge kann nach den einzelnen Satzteilen auf unterschiedliche Art gefragt werden (z. B. durch grammatische Transformation der Aussage in einem Fragesatz oder in Form eines Lückentests). Die Antwort kann frei zu formulieren sein oder unter mehreren Alternativen auszuwählen, usw. Die Schwierigkeit eines Tests hängt aber von solchen willkürlich festgesetzten Entscheidungen stark ab. Wenn nun zum selben Lehrziel mit gleichguten Gründen recht unterschiedliche Tests als inhaltlich valide zusammengestellt werden können, so stellt sich erneut die Frage nach der Übereinstimmung dieser Tests und der Äquivalenz der für das Erreichen des Lehrziels gestellten Anforderungen. Versucht man diese Frage abzuschneiden, indem man wie Klauer (1987) definiert "Lehrziel ist eine Aufgabenmenge", verschiebt man das Problem nur auf die Frage des Zusammenhangs zwischen Lehrzielen, die sich auf denselben Lehrstoff beziehen.

Weiter sollte nicht übersehen werden, daß auch Aufgaben, die aus demselben Aufgabenschema generiert wurden, weder gleich schwierig noch sonstwie psychologisch gleichwertig zu sein brauchen. Die Multiplikationsaufgaben "2 x 2 = ?" und "7 x 8 = ?" sind zwar aus demselben Schema generiert, die erste ist aber augenscheinlich leichter. In diesem Punkt führen die Arbeiten von Scandura (1977) weiter. Er faßt Lehrziele als Probleme auf, die mit einem bestimmten Algorithmus gelöst werden können. So z. B. geben Dumin & Scandura (1977) einen mehrstufigen Lösungsalgorithmus zur Subtraktion an. Aufgaben sind äquivalent, wenn zu ihrer Lösung der Lösungsalgorithmus in derselben Weise zu durchlaufen ist. Damit ist zumindest der Versuch gemacht, bei der Festlegung der Aufgaben, die ein Lehrziel repräsentieren, die am Lösungsprozeß beteiligten kognitiven Prozesse zu berücksichtigen. Dabei dürfte freilich der Bereich, in dem sich Lernen als Erwerb von Algorithmen auffassen läßt, inhaltlich begrenzt sein.

Die wichtigsten Ansätze zur Beantwortung der Frage, wie man inhaltliche Validität erreicht, und die mit den einzelnen Lösungsvorschlägen verbundenen Probleme konnten hier nur in den Grundzügen dargestellt werden. Übersichtstabelle 6.1 enthält eine Zusammenfassung. Detailliertere Darstellungen und Literaturhinweise zu den einzelnen Verfahren findet man unter anderem bei Roid & Haladyna (1982), Wieberg (1983) und Klauer (1987).

Wenn es um die praktische Anwendung der einzelnen Verfahren im Unterrichtsalltag geht, sind neben theoretischen Gesichtspunkten auch Fragen der Ökonomie zu bedenken. Speziell dann, wenn ein umfangreicher Lehrstoff, z.B. das in einem Schuljahr zu vermittelnde Geschichtswissen, mit formalsprachlichen Methoden in eine Aufgabenmenge umgesetzt werden soll, dürfte zumindest für Lehrer die Grenze des Zumutbaren überschritten sein.

Übersicht 6.1: Ansätze zur Konstruktion inhaltsvalider Tests

(a) Umgangssprachlich orientierte Anleitungen zur Umsetzung von Lernzielen in lernzielorientierte Tests. Beispiele:
Lehrzieloperationalisierung nach Mager (1965). Es ist zu operationalisieren, (1) welches Verhalten (2) unter welchen Bedingungen gezeigt werden soll und (3) bei welchem Kriterium des Lernziel als erreicht gilt.
Erstellen einer Tyler-Matrix (Tyler 1950, Bloom 1971). Unterscheidung von Inhaltsaspekt (z.B. Teilabschnitte des Lehrstoffs) und Handlungsaspekt (Wissen, Verstehen, Anwendung). Bei der Konstruktion der Testaufgaben werden diese beiden Aspekte systematisch miteinander kombiniert.
Vorteil: Breite Anwendbarkeit auf nahezu beliebige Bereiche; Nachteil: Relativ breiter Ermessensspielraum des Anwenders, so daß verschiedene Testkonstrukteure zu recht unterschiedlichen Tests kommen können.

(b) Formalsprachlich orientierte Verfahren zur Abgrenzung von Aufgabenuniversa Beispiele:
Osburn (1968), Hively et. al. (1968): Ein Aufgabenuniversum wird durch ein formales Aufgabenschema definiert, aus dem durch Einsetzen aus einer eindeutig definierten Menge von Zahlen, Begriffen usw. Aufgaben entstehen.
Scandura (1977): Ein Aufgabenuniversum wird dadurch definiert, daß sich die Aufgaben mit demselben Algorithmus lösen lassen.
Klauer (1987): Ein Lehrziel ist eine Aussagenmenge (z.B. Aussagen über den Dreißigjährigen Krieg). Daraus werden nach einer Sampling-Vorschrift (z.B. nach dem Zufall) Aussagen gezogen und nach bestimmten Transformationsregeln (z.B. Fragen nach einzelnen Satzteilen) in Testaufgaben umgeformt.
Vorteil: Das Aufgabenuniversum ist klar abgegrenzt. Nachteile: (1) Bei der Konstruktion des Aufgabenuniversums muß eine Vielzahl von Ad-hoc-Entscheidungen getroffen werden, so daß auch hier zum selben Lehrstoff recht unterschiedliche Tests entstehen können. (2) Bei umfangreicherem Lehrstoff aus ökonomischen Gründen schwer zu realisieren.

6.2.3.2 Das Binomialmodell und darauf aufbauende Klassifikationsstrategien

Wenn es gelungen ist, für ein Lehrziel einen inhaltsvaliden Test zu konstruieren, so stellt sich für eine kriteriumsorientierte Testinterpretation als nächstes die Frage, ab wann das Lehrziel als erreicht gelten soll und mit welcher Sicherheit im Einzelfall darüber entschieden werden kann. Solche Fragen wurden vor allem auf der Grundlage des Binomialmodells diskutiert. Man geht dabei von der (bislang allerdings kaum praktisch realisierten) Vorstellung aus, daß ein Aufgabenuniversum definiert ist, aus dem zufällig gezogen werden kann. Das Binomialmodell setzt voraus, daß jeder Proband v durch eine im Laufe der Testdurchführung gleichbleibende Wahrscheinlichkeit π_v (π = griechisch: pi) charakterisiert ist, mit der er ein Item aus dem Itempool löst. Die Wahrscheinlichkeit, daß der Proband von n vorgelegten Items k richtig löst, ergibt sich dann gemäß der Binomialverteilung. Der Erwartungswert dieser Bino-

mialverteilung ist der wahre Wert des Probanden $\tau_v = n\pi_v$, die Varianz die Fehlervarianz $\sigma^2(F_v) = n\pi_v(1-x")$. Im Unterschied zu den Annahmen, die gewöhnlich im Rahmen der klassischen Testtheorie bei der Berechnung von Konfidenzintervallen gemacht werden (Normalverteilung der Fehler, gleiche Fehlervarianz in allen Skalenbereichen; vgl. Kapitel 2.2), hängt hier die Fehlervarianz vom wahren Wert ab (geringere Fehlervarianz bei sehr kleinen oder sehr großen Werten von x_v; größte Fehlervarianz bei $\pi_v = 0.5$). Die Fehlerverteilung ist keine Normalverteilung, sondern eine um eine Konstante verschobene Binomialverteilung und entsprechend dem Wert von π_v rechts- oder linksschief.

Auf diesen Modellannahmen aufbauend, können Entscheidungsregeln entwickelt werden, wie diagnostiziert werden soll, ob ein Proband das Lehrziel erreicht hat: Dazu wird zunächst eine bestimmte Lösungswahrscheinlichkeit π_{krit} als für das Erreichen des Lehrziels entscheidend definiert. Je nachdem, wie man Fehler der einen oder anderen Art (ein "Könner" wird fälschlich als "Nichtkönner" oder umgekehrt ein "Nichtkönner" fälschlich als "Könner" eingestuft) gewichtet, und je nach Länge des Tests, kann die im Test erforderliche Trefferzahl nach oben oder unten wandern. In weiteren Varianten können Fehler, je nach Abstand des Probanden von der Kriteriumsmarke, unterschiedlich gewichtet werden, es können statt zwei auch drei kritische Punktwerte (Könner, unentschieden, Nichtkönner) festgesetzt werden, es können Annahmen über die Verteilung der wahren Werte mitberücksichtigt werden, usw. In Zensierungsmodellen wird für jede Note eine Lösungswahrscheinlichkeit als Kriteriumsmarke definiert, woran sich dann analoge Überlegungen anschließen lassen. Eine ausführliche Darstellung findet man bei Klauer (1987).

Das einfache Binomialmodell setzt voraus, daß für einen Probanden die Trefferwahrscheinlichkeit bei allen Items eines Tests gleich ist. Das ist erfüllt, wenn alle Items gleich schwierig sind - was praktisch unrealistisch ist. Wie bereits erwähnt, brauchen Items, die aus dem gleichen Aufgabenschema erzeugt sind, keineswegs gleich schwierig sein. Eine weitere Möglichkeit, das Modell zu realisieren besteht darin, aus einem beliebigen Itempool für jede Testdurchführung unabhängig Items zu ziehen. Auch wenn die Itemschwierigkeiten sehr unterschiedlich sind (die Versuchsperson z. B. 70 % der Vokabeln weiß und mit der Wahrscheinlichkeit 1 löst, die übrigen mit der Wahrscheinlichkeit 0), so ergibt sich durch die zufällige Itemauswahl eine konstante Trefferwahrscheinlichkeit ($\pi_v = 0{,}7$). Beispiel 6.1 geht von dieser Modellannahme aus. Verallgemeinerungen des Modells, bei denen die Voraussetzung einer konstanten Trefferwahrscheinlichkeit fallen gelassen wird und mit verschiedenen Näherungsverfahren ähnliche Fragen behandelt werden wie für das einfache Binomialmodell, findet man bei Klauer (1987).

Versucht man das Binomialmodell mit anderen testtheoretischen Modellen in Beziehung zu setzen, so läßt es sich wie folgt einordnen: Es läßt sich im Rahmen der klassischen Testtheorie als Spezialfall betrachten, der durch bestimmte Annahmen über die Meßfehlerverteilung charakterisiert ist. Geht man von der (wie bereits gesagt, wenig realistischen) Annahme gleicher Aufgabenschwierigkeit für alle Items aus, so läßt sich das Binomialmodell als Spezialfall eines Latent-Trait-Modells mit für alle Items gleicher Itemcharakteristik (z.B. Rasch-Modell mit für alle Items gleichem Schwierigkeitsparameter) betrachten.

Betrachtet man den Fall ungleicher Itemschwierigkeiten, bei dem die gleichbleibende Trefferwahrscheinlichkeit durch die zufällige Itemauswahl realisiert wird, so kann man zwischen Testwiederholung mit demselben Test (d. h. denselben Testitems)

Beispiel 6.1 Berechnung der Trefferverteilung bei gegebener Lösungswahrscheinlichkeit im Binomialmodell

Ein Prüfer hat festgesetzt, daß die Vokabelprüfung bestanden ist, wenn der Proband 80% der geprüften Vokabeln gewußt hat. Die Vokabeln werden zufällig gezogen. Ein Proband hat 90% der Vokabeln gelernt. Wie sicher kann er sein, daß er die Prüfung bestehen wird,

(a) wenn 10 Vokabeln geprüft werden, von denen 8 richtig sein müssen?

(b) wenn 20 Vokabeln geprüft werden, von denen 16 richtig sein müssen?

Lösung:

Einer Tabelle für die Binomialverteilung (z.B. Bortz, 1989) entnimmt man, daß bei n = 10 Versuchen und einer Fehlerwahrscheinlichkeit von $1 - \pi = 0.1$ die Wahrscheinlichkeit für 0 bis 1 Fehler (und damit 8 oder mehr Richtige) 0.73 beträgt. Bei n = 20 Versuchen ist die Wahrscheinlichkeit für 0 bis 4 Fehler (und 16 oder mehr Richtige) 0.96. D.h., bei einem längeren Test kann der Proband, der ja seinem Wissensstand nach über der 80%-Marke liegt, sicherer sein, die Prüfung tatsächlich zu bestehen.

und einem neu aus dem Itempool gezogenen Test unterscheiden. Solche Überlegungen führen dann zur Theorie der Generalisierbarkeit, die sich auf "zufallsparallele" Tests als Mitglieder einer "Testfamilie" gut anwenden läßt (siehe Kapitel 6.1).

Zusammenfassung

Versucht man nun, mehr als 20 Jahre nach dem Beginn der Diskussion um normorientierte versus kriterienorientierte Messung eine Bilanz zu ziehen, so ergibt sich folgendes Bild: Der in der frühen Diskussion betonte Gegensatz ist überwunden. Sowohl der Vergleich mit einer Normpopulation als auch der Vergleich mit einem inhaltlich definierten Standard als Kriterium ist diagnostisch relevant, und es hängt von der individuellen Problemstellung ab, ob eher der eine oder andere Gesichtspunkt im Vordergrund steht. Normorientierte und kriterienorientierte Messung setzen auch nicht zwei verschiedene Klassen von Tests voraus, sondern beide Arten von Informationen können aus demselben Test gewonnen werden, vorausgesetzt, daß er inhaltsvalide ist und Normdaten vorliegen.

Kernpunkt einer speziellen Testtheorie lehrzielorientierter Tests ist die Frage, wie inhaltliche Validität zu erreichen ist. Von unterschiedlichen Ansätzen ausgehend wurden Regelsysteme entwickelt, nach denen ein Lehrstoff in eine Aufgabenmenge umzusetzen ist. Solche Anleitungen zur Konstruktion von Testaufgaben können zweifellos sowohl zum Erstellen informeller Tests (z. B. Klausuren) als auch für die Testkonstruktion im engeren Sinn von Nutzen sein. Trotzdem bleiben Probleme bestehen:

(1) Die Umsetzung eines Lehrstoffs in eine Aufgabenmenge erfordert eine Reihe subjektiver Entscheidungen, so daß aus demselben Lehrstoff recht unterschiedliche Aufgabenuniversa erzeugt werden können.

(2) Auch eine eindeutige Definition eines Aufgabenuniversums - z. B. mit Hilfe einer generierenden Regel - impliziert keineswegs die psychologische Homogenität der generierten Items.

Weitere spezifische Beiträge, die durch die speziellen Probleme kriterienorientierter Messung initiiert wurden, sind die Weiterentwicklung des Binomialmodells und die darauf aufbauenden Klassifikationsstrategien und Zensierungsmodelle.

Einführende Literatur:

Strittmatter, P. (Hrsg.) (1973). ***Lehrzielorientierte Leistungsmessung.*** Weinheim: Beltz.

Weiterführende Literatur:

Fricke, R. (1974). ***Kriteriumsorientierte Leistungsmessung.*** Stuttgart: Kohlhammer.

Klauer, K.J. (1987). ***Kriteriumsorientierte Tests.*** Göttingen: Hogrefe.

Stelzl, I. (1976). Versagt die klassische Testtheorie bei kriterienorientierten Tests? ***Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 8,*** 106-116.

Tent, L. & Waldow, M. (1984). Pädagogische Diagnostik in der Schule für Lernbehinderte: Gruppenbezogene Leistungsmessung oder Zielerreichungs-Tests? ***Heilpädagogische Forschung,*** 11, 1-29.

Wieberg, H.-J.W. (1983). Probleme kriteriumsorientierter Leistungsmessung. In R. Horn, K. Ingenkamp & R.S. Jäger (Hrsg.), ***Tests und Trends*** 3. Weinheim, Beltz.

6.3 Methodische Beiträge zum Problem der Testfairness

1. **Wann ist ein Test gegenüber allen gesellschaftlichen Gruppen als "fair" zu bezeichnen?**
2. **Welche Versuche wurden gemacht, den Begriff der Testfairness von der methodischen Seite her zu definieren, und wo stoßen diese Bemühungen auf Grenzen?**

Vorstrukturierende Lesehilfe

Das Problem der Testfairness wurde zunächst vor allem in den USA, dort hauptsächlich in Zusammenhang mit Fragen der Diskriminierung rassischer Minderheiten, intensiv diskutiert. Nicht zuletzt durch die Klage eines weißen Amerikaners gegen ein Universitäts-Zulassungverfahren, das rassischen Minderheiten einen Bonus einräumte (dpa-Meldung September 1977, zitiert nach Möbus, 1978), wurde die Diskussion um Chancengleichheit und Testfairness weiter angeheizt.

Im deutschen Sprachraum waren es nicht zuletzt die bildungspolitischen Probleme im Zusammenhang mit Hochschulzulassungsverfahren (Tests für die medizinischen Studiengänge), die zu einer öffentlichen Diskussion um die Frage möglicher Benachteiligung bestimmter Personengruppen, z.B. Angehöriger unterer sozialer Schichten, führten. Differenziertere Konzepte der Testfairness, die in den USA bereits entwickelt waren, wurden aufgegriffen und weiter diskutiert. Im folgenden wird zunächst das prognose-orientierte Testfairness-Konzept vorgestellt. Danach ist die Selektion mithilfe eines Tests z.B. gegenüber Angehörigen aller sozialen Schichten fair, wenn in allen sozialen Schichten der gleiche Testwert der gleichen Erfolgswahrscheinlichkeit entspricht (6.3.1). Dieses Konzept hilft allerdings nicht weiter, wenn es darum geht, festzulegen, welche Merkmale trotz möglicher prognostischer Relevanz grundsätzlich nicht zur Prognose herangezogen werden sollen, weil das offensichtlich unbillig wäre (6.3.2). Radikale Alternativen zum prognose-orientierten Testfairness-Konzept sind das Identitätskonzept und Quotenpläne.

6.3.1 Das prognose-orientierte Testfairness-Konzept

Dort, wo Tests zur Selektion von Bewerbern eingesetzt werden, wie z.B. bei Hochschulzulassungsverfahren, werden Probleme der Testfairness am deutlichsten sichtbar: Die Tests sollen allen Bewerbern die gleiche Chance geben, es sollen nicht einzelne Gruppen (z.B. nach dem Geschlecht, nach der sozialen Herkunft o.ä.) bevorzugt oder benachteiligt werden. Diese Fragen sind natürlich besonders brisant, wenn Entscheidungen auch gegen den Willen der Betroffenen (Ablehnung von Bewerbern) von öffentlichen Institutionen getroffen werden. Sie stellen sich grundsätzlich aber auch in Beratungssituationen, bei denen die Entscheidung vom Ratsuchenden selbst getroffen wird: Eine systematische Fehleinschätzung bestimmter Personengruppen auf-

grund mangelnder Testfairness würde sich zwar nicht so unmittelbar, aber der Tendenz nach ähnlich auswirken wie bei direkter Selektion.

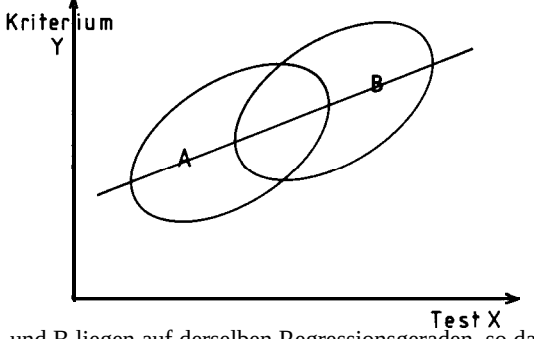
So einfach und einleuchtend die Forderung nach fairen Test- und Selektionsverfahren zunächst auch aussieht, so zeigt sich bei näherem Hinsehen doch bald, daß es sich beim Begriff der Testfairness um ein theoretisch schwieriges Konzept handelt, das in verschiedener Weise expliziert werden kann. Versucht man zunächst die Forderung, ein faires Selektionsverfahren müsse jedem die gleiche Chance einräumen, ganz wörtlich zu nehmen, so ist sie am besten durch das Los zu erfüllen. Daß lebenslaufbestimmende Entscheidungen völlig grundlos, unvorhersehbar und unbeeinflußbar per Zufall getroffen werden, dürfte indes kaum jemand für wünschenswert halten. Wenn von Tests verlangt wird, sie sollten jedem die gleiche Chance geben, so ist damit sicher nicht gemeint, sie sollten nach dem Zufall funktionieren, sondern es ist etwas mitgedacht, was nicht explizit gemacht ist: Jedem die gleiche Chance bei gleichem Leistungsstand, bei gleichen Fähigkeiten, gleicher Erfolgswahrscheinlichkeit gemessen an verschiedenen Bewährungskriterien. Vom Test ist dann zu fordern, daß er ein möglichst valider Indikator für diejenigen Kriterien (Ausbildungs- und Berufserfolg) ist, die die Selektion bestimmen sollen. Von daher erscheint die Test-Kriteriumsbeziehung ein geeigneter Ansatzpunkt, um Testfairness begrifflich näher zu bestimmen. Eine Reihe von Autoren hat die Test-Kriteriumsbeziehung als Ausgangspunkt gewählt, um ein statistisches Konzept der Testfairness zu entwickeln, das im folgenden dargestellt werden soll.

Nach Cleary (1968) und Anastasi (1968) ist ein Test X zur Vorhersage eines Kriteriums Y (z.B. Studienerfolg) fair gegenüber den Gruppen i=l ... g (z.B. Bewerbern aus verschiedenen Schularten), wenn für alle Gruppen dieselbe Test-Kriteriumsbeziehung gilt, so daß bei allen Gruppen gleicher Testleistung gleiche durchschnittliche Kriteriumsleistungen entsprechen. Im Falle einer linearen Test-Kriteriumsbeziehung bedeutet das, daß dieselbe Regressionsgerade

$$E(Y/x) = a+ß\ x$$

mit denselben Werten für α und ß für alle g Gruppen gültig ist. Eine solche Situation ist in Abbildung 6.1 dargestellt:

Abbildung 6.1: Test-Kriteriums-Beziehung, die die Testfairness-Bedingung nach Cleary (1968) & Anastasi (1968) erfüllt.

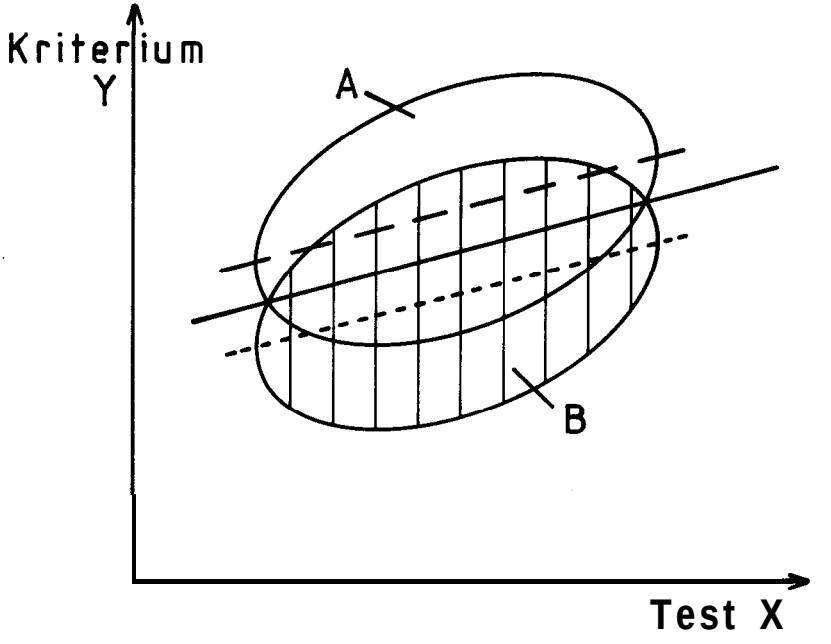


Die Populationen A und B liegen auf derselben Regressionsgeraden, so daß demselben Testwert X in beiden Populationen derselbe vorhergesagte Kriteriumswert Y* = E(Y/x) entspricht.

Die beiden Gruppen A und B unterscheiden sich sowohl hinsichtlich der durchschnittlichen Testleistung als auch hinsichtlich der durchschnittlichen Kriteriumsleistung, es gilt aber für beide Gruppen dieselbe Regressionsgerade. Für beide Gruppen ist der durchschnittliche Schätzfehler (Abweichung des tatsächlichen Y-Werts vom Regressionsschätzwert) gleich Null, d.h. bei Verwendung des Tests als Prädiktor wird für keine der beiden Gruppen der Kriteriumswert systematisch über- oder unterschätzt. In Abbildung 6.1 sieht man das daran, daß bei jeder der beiden Gruppen der gleiche Flächenanteil über der Regressionslinie liegt (die tatsächlichen Kriteriumswerte sind höher als die Regressionsschätzung, der Kriteriumswert des Pbdn wird also unterschätzt) wie unterhalb der Regressionslinie (der tatsächliche Kriteriumswert ist niedriger als die Regressionsschätzung, der Kriteriumswert des Pbdn wird überschätzt).

Abbildung 6.2a und 6.2b zeigen Fälle, in denen die von Cleary (1968) und Anastasi (1968) angegebene Bedingung nicht erfüllt ist.

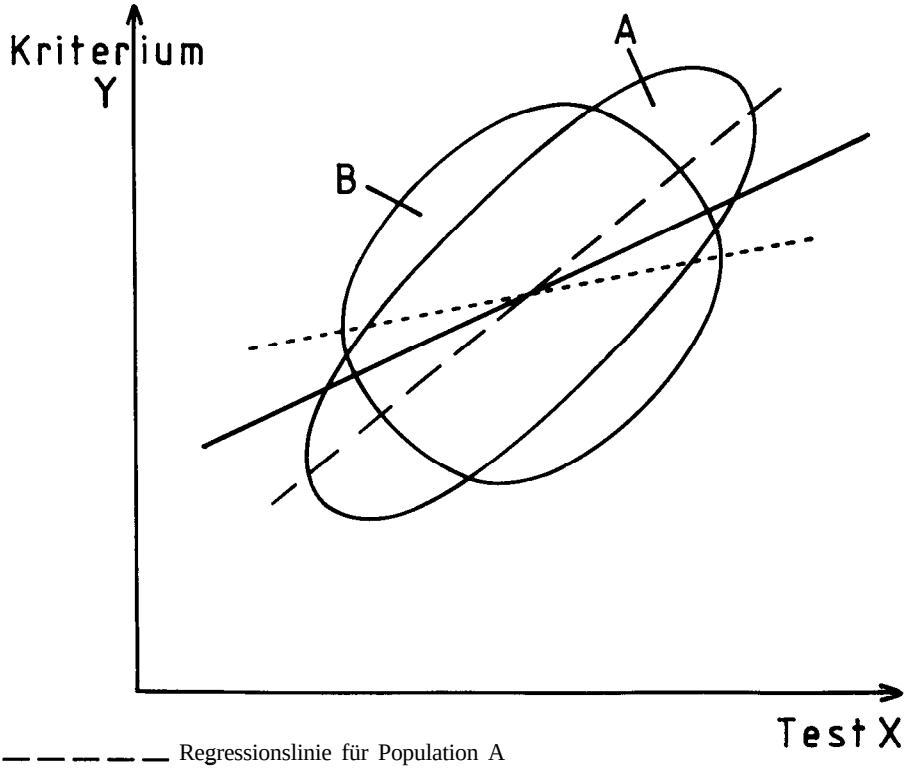
Abbildung 6.2a: Mangelnde Testfairness im Sinne von Cleary (1968) & Anastasi (1968): Unterschied im Mittelwert des Kriteriums bei sonst gleicher Test-Kriteriums- Beziehung.



– – – – – Regressionslinie für Population A
..... Regressionslinie für Population B
——— Gemeinsame Regressionslinie bei Zusammenfassung beider Populationen

Die Kriteriumswerte der Probanden aus Population A liegen häufiger über der gemeinsamen Regressionslinie, d. h. die Probanden werden bei Verwendung der gemeinsamen Regressionslinie unterschätzt. Bei Population B ist es umgekehrt.

Abbildung 6.2b: Mangelnde Testfairness im Sinn von Cleary (1968) & Anastasi (1968): Unterschiedliche Steigung der Regressionsgeraden bei gleichen Mittelwerten.



— — — — — Regressionslinie für Population A
..... Regressionslinie für Population B
——————— Gemeinsame Regressionslinie bei Zusammenfassung beider Populationen

Bei Verwendung dergemeinsamen Regressionslinie werden in Population A die Kriteriumswerte von Probanden mit hohen Testwerten im Durchschnitt unterschätzt, die von Probanden mit niedrigen Testwerten im Durchschnitt überschätzt. In Population B ist es umgekehrt.

In Abbildung 6.2a unterscheiden sich die beiden Gruppen nicht im durchschnittlichen Testwert, wohl aber in den durchschnittlichen Kriteriumswerten. Bei Verwendung einer aus beiden Gruppen gemeinsam bestimmten Regressionsgleichung werden die Kriteriumswerte der Gruppe A unterschätzt, die der Gruppe 13 überschätzt. Dem Unterschied im durchschnittlichen Kriteriumswert entspricht in der Regressionsgleichung ein Unterschied in der Regressionskonstanten.

Durch die Verwendung gruppenspezifischer Regressionsgleichungen wird hier die Güte der Vorhersage verbessert, d.h. die Schätzfehler werden dem Betrag nach verringert. Außerdem sind dann in jeder der beiden Gruppen positive und negative Schätzfehler gleich häufig, so daß der durchschnittliche Fehler in jeder Gruppe Null ist. Im vorliegenden Spezialfall (lineare Regression, gleicher Anstieg der Regressi-

onsgeraden) läuft die Berechnung getrennter Regressionslinien auf einen Bonus für die Gruppe A hinaus, der dem Unterschied in der Regressionskonstanten entspricht. Ein solcher Bonus für Angehörige der Gruppe A entspricht in anderer formaler Darstellung der Verwendung einer multiplen Regressionsgleichung mit "Gruppenzugehörigkeit" als zweitem, zum Test hinzukommendem Prädiktor.

In Abbildung 6.2b ist eine Situation dargestellt, in der beide Gruppen sowohl im Test als auch im Kriterium gleiche Mittelwerte haben. Die Test-Kriteriumskorrelation ist aber bei Gruppe A höher und der Anstieg der Regressionslinie daher steiler als bei Gruppe B. Bei Verwendung einer aus beiden Gruppen gemeinsam berechneten Regressionslinie ist zwar für jede der beiden Gruppen der durchschnittliche Schätzfehler Null (es liegen gleich viele Probanden über und unter der gemeinsamen Regressionsgeraden), trotzdem gibt es systematische Tendenzen: Bei den Probanden der Gruppe A werden bei überdurchschnittlichen Testleistungen im Durchschnitt zu niedrige, bei unterdurchschnittlichen Testleistungen im Durchschnitt zu hohe Kriteriumsleistungen vorhergesagt. Bei Probanden der Gruppe B verhält sich das genau umgekehrt. Auch hier würde eine Berechnung getrennter Regressionslinien die Kriteriumsvorhersage insgesamt verbessern. Da hier der Unterschied im Anstieg der Regressionsgeraden liegt, ist die Berechnung getrennter Regressionslinien nicht als additiver Zuschlag (Bonus) für eine der beiden Gruppen darstellbar.

Man beachte, daß gemäß der Definition von Cleary (1968) und Anastasi (1968) Testfairness ein Begriff ist, der eine dreistellige Relation aus Test, Gruppenzugehörigkeit und Kriterium beinhaltet. Die bloße Anwendung eines Tests bei einer Gruppe, ohne Bezugnahme auf ein Kriterium, ist demnach noch nicht als fair oder unfair zu beurteilen. So wäre die Feststellung, daß die Schüler der Schule A, die nur zwei Jahre Physikunterricht hatten, in einem Physiktest schlechter abgeschnitten haben als die Schüler der Schule B, die vier Jahre Physik hatten, nicht unfair. Würde dieser Test jedoch verwendet, um die Studieneignung für Medizin vorherzusagen, so würden sich die Schüler der Schule A wahrscheinlich zu Recht über Unfairness beklagen, wenn für alle Schüler dieselbe Regressionsgleichung bzw. derselbe kritische Punktwert verwendet würde. Bei Aufteilung der Daten nach Schulart würde sich voraussichtlich zeigen, daß verschiedene Regressionsgleichungen gelten und bei Verwendung einer gemeinsamen Regressionsgleichung die Kriteriumswerte der Gruppe A systematisch unterschätzt werden. Bei der Verwendung getrennt berechneter Regressionsgleichungen wäre dann innerhalb jeder Gruppe eine höhere Korrelation zwischen Test und Kriterium zu finden als in der aus beiden Schularten gemischt zusammengesetzten Population, so daß die Verwendung getrennter Regressionsgleichungen auch insgesamt eine bessere Vorhersagegenauigkeit ergäbe als die Verwendung einer gemeinsamen Regressionslinie.

Auch Fälle, in denen beim ersten Hinsehen kein Bezug auf ein Kriterium zu erkennen ist, lassen sich im Rahmen des Prognose-orientierten Testfairness-Konzepts interpretieren. Wenn die Verwendung eines verbalen Intelligenztests bei fremdsprachigen Ausländern zu Recht als unfair bezeichnet würde, so deshalb, weil damit eine Generalisierung von spezifischen Unkenntnissen auf andere Leistungsbereiche nahegelegt würde. Die Kriterien werden zwar nicht ausdrücklich genannt, aber es würde bezüglich eines breiten Feldes möglicher Kriterien eine Unterschätzung erfolgen.

Weiter kann ein Test X, der sich bei der Vorhersage eines Kriteriums Y bezüglich einer bestimmten Gruppenaufteilung (z.B. Schulart) als fair erwiesen hat, sich bei Verwendung anderer Gruppierungsmerkmale (z.B. Geschlecht) als unfair erweisen.

Wottawa & Amelang (1980) weisen zu Recht darauf hin, daß sich bei jeder Kriteriumsvorhersage eine Vielzahl von Gruppierungsvariablen finden läßt, die mit dem Schätzfehler korrelieren und deren Hinzunahme als Prädiktoren die Kriteriumsvorhersage verbessern würde. Wenn aus einem Intelligenztest X die Schulleistung Y vorhergesagt wird, so würde vermutlich der Fleiß als zusätzlicher Prädiktor eine Verbesserung der Kriteriumsvorhersage erbringen. Das heißt aber nichts anderes, als daß bei Verwendung des Tests allein die Kriteriumswerte der Fleißigen systematisch unterschätzt, die der Faulen überschätzt werden. Im Sinne der Definition von Cleary (1968) & Anastasi (1968) bedeutet das mangelnde Testfairness bzw. Selektionsfairness gegenüber den Fleißigen. Da es nun sicher überzogen wäre, zu sagen ein Test wäre nur dann fair einsetzbar, wenn die Vorhersage durch keine weiteren Prädiktoren verbesserungsfähig ist, wird man sich entscheiden müssen, bezüglich welcher Merkmale Testfairness untersucht und nötigenfalls durch Verwendung entsprechend modifizierter Selektionsstrategien (Bonus/Malus-System, Berechnung getrennter Regressionsgleichungen) hergestellt werden soll.

Das zunächst von Cleary (1968) & Anastasi (1968) vorgestellte Konzept der Testfairness wurde von verschiedenen Autoren weiter diskutiert und modifiziert. So z.B. gehen Einhorn & Bass (1971) von der Vorstellung eines kritischen Kriteriumswertes aus, ab dem jemand als erfolgreich gelten soll (z.B. Bestehen der Abschlußprüfung mit mindestens "ausreichend"), und fordern, daß der kritische Testwert gruppenspezifisch jeweils so festgelegt wird, daß dem Erreichen dieses Testwerts dieselbe Erfolgswahrscheinlichkeit entspricht.

$$p(Y>y_{krit}/X=x_i^*) = \text{konstant für alle Gruppen i}$$

y_{krit}= Kriteriumswert, ab dem jemand als erfolgreich gilt

X_i^*= gruppenspezifisch festgelegter, für eine Aufnahme erforderlicher Testwert.

Das Modell fordert also eine für alle Gruppen gleiche minimale Erfolgswahrscheinlichkeit, ab der eine Aufnahme erfolgt, ohne daß eine bestimmte Form der Test-Kriteriumsbeziehung (z.B. lineare Regression) zugrunde gelegt wird.

Cole (1973) und Linn (1973) gehen ebenfalls von einem dichotomen Erfolgskriterium aus. Cole (1973) fordert, daß bei allen Gruppen die Wahrscheinlichkeit für einen Bewerber, aufgenommen zu werden, wenn er geeignet ist, gleich sein soll. Der für eine Aufnahme erforderliche Testwert soll dementsprechend gruppenspezifisch festgelegt werden. Bei stark ungleichen Grundquoten (=Anteilen an Geeigneten) in den einzelnen Gruppen führt eine solche Selektion zu entsprechend ungleichen Anteilen ungeeignet Aufgenommener aus den verschiedenen Gruppen, was aber bewußt in Kauf genommen wird.

Linn (1973) wiederum schlägt vor, die für die Aufnahme erforderlichen Testwerte so festzulegen, daß bei allen Gruppen der Anteil der Erfolgreichen an den Aufgenommenen gleich ist, wobei dann der Anteil der abgelehnten Geeigneten ungleich sein kann.

Weitere Varianten des prognose-orientierten Konzepts der Testfairness sollen hier nicht dargestellt werden. Eine Übersicht findet man bei Möbus (1978).

6.3.2 Probleme des prognose-orientierten Testfairness-Konzepts

In der folgenden Diskussion greifen wir der Einfachheit halber auf das Regressionskonzept von Cleary (1968) & Anastasi (1968) zurück, wobei die Argumentation für verwandte Konzepte analog zu führen wäre.

Nach Cleary (1968) & Anastasi (1968) ist eine Testanwendung fair, wenn bei allen Probandengruppen gleichen Testwerten gleiche durchschnittliche Kriteriumswerte entsprechen. Statt eines einzelnen Tests kann natürlich auch eine Testbatterie oder ein aus Testdaten und anderen Informationsarten zusammengesetzter Wert verstanden werden. Statt von Testfairness ist dann von Fairness der Selektionsstrategie zu sprechen.

Bei den Kriterien, die vorhergesagt werden sollen, geht es gewöhnlich um Ausbildungserfolg (Erreichen des Abschlusses, Noten, Beurteilungen durch Lehrer und Ausbilder) und berufliche Bewährung (Ausüben des erlernten Berufs, Zufriedenheit, Selbst- und Fremdbeurteilung des Erfolgs). Bei der Vorhersage solcher Kriterien werden sich allerdings eine Reihe von Merkmalen als gute Prädiktoren erweisen, bei denen man es als ausgesprochen unfair empfände, wenn sie zur Selektion herangezogen würden. Wenn z.B. ein Jugendlicher durch häusliche Umstände stark belastet ist (zerrüttete Familie, Belastung durch die Betreuung von Schwerkranken), wird das vermutlich den Ausbildungserfolg mindern. Die Kenntnis solcher Umstände zum Nachteil des Jugendlichen zu verwenden, würde wohl niemand als “fair” empfinden, egal ob damit die Vorhersage verbessert wird oder nicht. Andere Beispiele lassen sich leicht finden: Wer von den Eltern einen gut eingeführten Betrieb übernehmen kann, wird mit erhöhter Wahrscheinlichkeit in dem entsprechenden Beruf erfolgreich sein, wer eine Ausbildungseinrichtung wählt, in der die Leistungsanforderungen bekanntermaßen etwas geringer sind als üblich, wird mit erhöhter Wahrscheinlichkeit abschließen, usw.

An diesen Beispielen wird deutlich, daß eine formale Definition “gleiche Erfolgswahrscheinlichkeit = gleiche Selektionswahrscheinlichkeit” nicht ausreicht, um Selektionsfairness zu definieren, sondern daß es zusätzlich einer inhaltlichen Abgrenzung bedarf, auf welche Prädiktoren die Prognose zu stützen ist, und welche nicht herangezogen werden sollen. Letztere Frage ist nur unter Bezugnahme auf gesellschaftspolitische Wertsetzungen zu beantworten und geht damit über den Bereich empirischer Wissenschaft hinaus. Vermutlich werden die meisten einig sein, daß es fairer ist, die Prognose auf Eigenschaften des Probanden zu stützen (Fähigkeiten, Interessen, bisher erbrachte Leistungen) als aufäußere Umstände, die er nicht zu vertreten hat. Versucht man jedoch, beides zu trennen, so stößt man sehr rasch auf Abgrenzungsprobleme: Die Eigenschaften des Probanden sind Ergebnis einer Entwicklung, die seinen bisherigen Lebensbedingungen entspricht. Wenn jemand dank bestimmter Arbeitshaltungen und -techniken erfolgreicher studiert als andere, wird man ihm diesen Erfolg persönlich zuschreiben. Den Erwerb dieser Arbeitsweise hat er vielleicht einem engagierten Nachhilfelehrer zu verdanken. Die Frage, was dem Probanden selbst positiv oder negativ zuzurechnen ist und was nicht, und ob es fair ist, ein bestimmtes Merkmal zur Prognose heranzuziehen, wird deshalb in vielen Fällen strittig bleiben.

Diese inhaltlichen Schwierigkeiten machen es verständlich, daß dem prognoseorientierten Testfairness-Konzept simplere Konzepte gegenüberstehen, bei denen auf den Kriterienbezug und auf Validitätsmaximierung bewußt verzichtet wird.

6.3.3 Identitätskonzept und Quotenpläne als Alternativen zum prognoseorientierten Testfairness-Konzept

Das Identitätskonzept ist das einfachste Konzept der Testfairness. Danach ist ein Test fair gegenüber Probanden aus unterschiedlicher sozialer Schicht (mit unterschiedlicher Schulbildung, gegenüber beiden Geschlechtern usw.), wenn er keinerlei Zusammenhang mit der sozialen Schicht (bzw. dem in Frage stehenden Gruppierungsmerkmal) zeigt, d.h. die Testwerte müssen sich in allen sozialen Schichten (innerhalb jeder Schulart, bei beiden Geschlechtern usw.) gleich verteilen. Eine solche Forderung ist wohl vielfach nur zu erfüllen, wenn ganz erhebliche Abstriche vom inhaltlichen Testkonzept und damit von der Validität gemacht werden. Gerade wenn man davon ausgeht, daß z.B. die verschiedenen sozialen Schichten unterschiedlich günstige Entwicklungsbedingungen bieten, die sich in einem entsprechend unterschiedlichen Fähigkeitsstand niederschlagen, kann man von einem validen Fähigkeitstest nicht verlangen, daß er keine Abhängigkeit von der sozialen Schicht zeigt.

Wottawa & Amelang (1980) weisen darauf hin, daß das Identitätskonzept jederzeit auch ohne Eingriff in den Testinhalt realisiert werden kann, wenn man die Testwerte gruppenspezifisch normiert. In der Tat geben Test-Handanweisungen bisweilen mehrere Arten von gruppenspezifischen Normen an, z.B. Normen für verschiedene Altersstufen, für verschiedene Schulabschlüsse, getrennt nach Geschlechtern usw. Wendet man solche gruppenspezifische Normen, z.B. nach Schularten getrennte Normtabellen an, so korrelieren die normierten Werte ex definitione mit dem Gruppierungsmerkmal, hier dem erreichten Schulabschluß, zu Null. Eine Selektion aufgrund solcher gruppenspezifisch normierter Werte entspricht der Vergabe eines Bonus an Personen aus der Gruppe mit den niedrigeren Durchschnittswerten, da hier ja derselbe Normwert aufgrund eines niedrigeren Testrohwerts erreicht wird. Wenn es für die Vorhersage eines Kriteriums primär auf den tatsächlichen Leistungsstand ankommt, so wie er sich im Testrohwert ausdrückt, kann die Verwendung gruppenspezifischer Normwerte, im ganzen gesehen, nur zu einer Verschlechterung der Vorhersage führen. Letzteres in Kauf zu nehmen, braucht aber nicht irrational zu sein, wenn man unter der vorrangigen Zielsetzung, bestimmte gesellschaftliche Veränderungen durchzusetzen (z.B. Angehörige von Minderheiten verstärkt auf Hochschulen zu bringen), bewußt von einer Selektion nur nach dem zu erwartenden Erfolg absieht.

Noch deutlicher wird dieser Gesichtspunkt, wenn die Selektion nach einem Quotenplan erfolgt, so daß Angehörige verschiedener Gruppen nicht mehr gegeneinander konkurrieren, sondern jeder Gruppe ein bestimmtes Kontingent an Plätzen unabhängig von der Leistungsfähigkeit zugewiesen wird. Innerhalb jeder Gruppe kann dann wieder prognosenorientiert selegiert werden. Eine solche Quotierung vorab widerspricht der Zielsetzung, diejenigen Probanden auszuwählen, die bei den bestehenden gesellschaftlichen Bedingungen die höchste Erfolgserwartung haben. Das ist aber nicht als irrational zu betrachten, wenn dieses Ziel bewußt zurückgestellt wurde, z.B. in kompensatorischer Absicht oder in Hinblick auf erwartete Signalwirkungen auf andere gesellschaftliche Bereiche.

Zusammenfassung

Das Konzept “gleiche Erfolgswahrscheinlichkeit = gleiche Selektionswahrscheinlichkeit”, das den Grundgedanken des prognose-orientierten Testfairness-Konzepts ausmacht, stößt auf Grenzen, wenn aufgrund von Werthaltungen bestimmte Merkmale nicht zur Prognose herangezogen werden sollen. Da Erfolgswahrscheinlichkeit zumindest empirisch nur unter den jeweils gegenwärtigen gesellschaftlichen Bedingungen bestimmt werden kann, läuft eine Selektion nach Erfolgswahrscheinlichkeit Gefahr, bestehende Benachteiligungen/Vorteile zu reproduzieren. Wenn gesellschaftliche Veränderung das vorrangige Ziel ist, so kann statt einer Selektion nach Erfolgswahrscheinlichkeit eine Selektion nach Quotenplänen als zweckmäßiger erscheinen.

Einführende Literatur:

Möbus, C. (1978). Zur Fairness psychologischer Intelligenztests. Ein unlösbares Problem zwischen Gruppen, Individuen, Institutionen? *Diagnostica, 24,* 19 1-234.

Weiterführende Literatur:

Möbus, C. (1983). Die praktische Bedeutung der Testfairness als zusätzliches Kriterium zu Reliabilität und Validität. In: R. Horn, K. Ingenkamp & R.S. Jäger (Hrsg.), ***Tests und Trends*** 3 (S. 155-203). Weinheim: Beltz.

Wottawa, H. & Amelang, M. (1980). Einige Probleme der Testfairness und ihre Implikationen für Hochschulzulassungs-Verfahren. ***Diagnostica, 26,*** 199-221.

7. Latent-Trait-Modelle

1. Was sind die gemeinsamen Grundannahmen aller Latent-Trait-Modelle?
2. Welche speziellen Annahmen macht das Rasch-Modell, und welche spezifischen Vorzüge ergeben sich daraus?
3. Welche Weiterentwicklungen haben sich aus dem Rasch-Modell ergeben und wo liegen die wichtigsten Anwendungsbereiche für die Pädagogisch-psychologische Diagnostik?
4. Welche anderen probabilistischen Modelle gehen von ähnlichen Annahmen aus wie die Latent-Trait-Modelle?

Vorstrukturierende Lesehilfe

Zunächst werden die für alle Latent-Trait-Modelle grundlegenden Begriffe, nämlich der Begriff der Itemcharakteristik und der lokalen stochastischen Unabhängigkeit, eingeführt (7.1). Danach wird das Rasch-Modell, das aus dem allgemeinen Ansatz durch die Annahme logistischer Itemcharakteristiken hervorgeht, mit seinen speziellen Vorzügen (spezifische Objektivität, Existenz erschöpfender Statistiken) dargestellt (7.2). Der Ansatz des Rasch-Modells wurde in verschiedene Richtungen weiterentwickelt: Das linear-logistische Modell erlaubt es, Hypothesen über das Zustandekommen der Itemschwierigkeiten zu testen (7.3). Das mehrkategoriale Rasch-Modell läßt nicht nur zwei Antwortkategorien (richtig/falsch), sondern mehrere qualitativ oder quantitativ verschiedene Kategorien zu (7.4). Das zweiparametrige logistische Modell (Birnbaum-Modell) erweitert den Ansatz des Rasch-Modells um einen zusätzlichen Itemparameter, der Unterschiede in der Itemtrennschärfe ausdrückt, das dreiparametrige Modell fügt einen weiteren Parameter für die Ratewahrscheinlichkeit hinzu (7.5). Im letzten Abschnitt (7.6) wird auf andere probabilistische Modellansätze hingewiesen, die ebenfalls von der Annahme der lokalen Unabhängigkeit ausgehen und damit dem Latent-Trait-Ansatz nahestehen.

7.1 Der Latent-Trait-Ansatz

Während die klassische Testtheorie auf alle psychologischen Maße anwendbar ist (bei jedem Maß läßt sich die Frage nach Reliabilität und Validität stellen), machen Latent-Trait-Modelle mehr oder weniger restriktive Annahmen über das Zustandekommen eines Testwerts. Ziel ist es, den Test so zu konstruieren, daß er diesen Annahmen entspricht. Wenn das gelingt, ergeben sich daraus die aus dem entsprechenden Modell ableitbaren Vorzüge.

Allen Latent-Trait-Modellen gemeinsam ist die Annahme eines latenten Kontinuums (Fähigkeit, Eigenschaft) ξ (griechisch: ksi), auf dem jede Person v eine bestimmte Ausprägung 5_v aufweist. Die Wahrscheinlichkeit, daß eine Person v ein bestimmtes Item i löst, hängt von ihrem Wert auf dem latenten Kontinuum ab.

Im einfachsten Fall kann man annehmen, daß es für jedes Item einen kritischen Wert auf ξ gibt, ab dem die Aufgabe gelöst wird. Diese Annahme liegt dem Guttman-Modell zugrunde, einem deterministischen Modell, das als Vorläufer der später entwickelten probabilistischen Latent-Trait-Modelle anzusehen ist. Der Grundgedanke des Guttman-Modells läßt sich am einfachsten am Beispiel der Körpergröße illustrieren: Wir nehmen an, Personen würden in folgender Weise nach ihrer Körpergröße befragt: "Sind Sie größer als 150 cm?", "Sind Sie größer als 160 cm?" usw. Die Dimension ξ ist hier die wahre Körpergröße der Person. Die Wahrscheinlichkeit, daß ein Item mit "ja" beantwortet wird, springt jeweils an einer bestimmten Stelle von Null auf Eins (Das Item "Sind Sie größer als 150 cm?" wird von Personen bis unter 150 cm zu 0%, von Personen ab 150 cm zu 100% bejaht oder "gelöst"). Abbildung 7.1 zeigt eine Guttman-Skala mit drei Items unterschiedlicher Schwierigkeit.

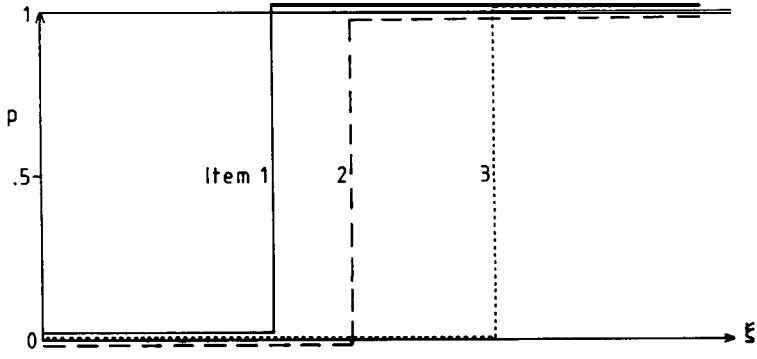


Abbildung 7.1: Guttman-Skala mit drei Items. Für jedes Item steigt an einer bestimmten Stelle des Merkmalskontinuums ξ die Lösungswahrscheinlichkeit p von Null auf Eins.

Wenn Items eine perfekte Guttman-Skala bilden, darf es nicht vorkommen, daß eine Person, die ein schwierigeres Item gelöst hat, ein leichteres verfehlt. Ordnet man die Items der Schwierigkeit nach aufsteigend an, so kann man der Angabe "der Proband hat k Aufgaben gelöst" zugleich entnehmen, welche Aufgaben er gelöst hat, nämlich alle Aufgaben mit Nummer 1 bis k und keine der Aufgaben ab Nummer k + 1.

Die Guttman-Skala ist zwar ein einfaches und zunächst plausibles Modell, doch ist bei psychologischen Daten kaum damit zu rechnen, daß es in dieser strikten Form erfüllt ist. Es kommt praktisch immer vor, daß Probanden ein leichtes Item, das sie bei ihrer Trefferzahl gelöst haben müßten, doch verfehlt haben, oder daß sie einzelne schwierigere Aufgaben überraschend doch lösen, nachdem sie mehrere leichtere nicht lösen konnten. Die Annahme, daß die Lösungswahrscheinlichkeit an einer bestimmten Stelle von Null auf Eins springt, ist sehr restriktiv und in der Testkonstruktion kaum zu erfüllen. Um anzugeben, inwieweit eine Guttman-Skala wenigstens annäherungsweise realisiert ist, wurden verschiedene ***Reproduzierbarkeitskoeffizienten*** (sie geben an, inwieweit aus den Trefferzahlen die genauen Antwortmuster "repro-

Beispiel 7.1: Lokale Unabhängigkeit bei festem ξ und Zustandekommen einer Itemkorrelation in einer in ξ variierenden Population.

Wir nehmen an, eine Person mit der Fähigkeitsausprägung ξ_1 löse Item i mit der Wahrscheinlichkeit $p(i+/\xi_1) = 0.1$ und Item j mit $p(j+/\xi_1) = 0.3$. Für eine andere Person mit der Fähigkeitsausprägung ξ_2 seien die entsprechenden Lösungswahrscheinlichkeiten $p(i+/\xi_2) = 0.7$ und $p(j+/\xi_2) = 0.9$. Nimmt man für jede Person an, daß die Itembeantwortung unabhängig erfolgt, so ergeben sich die als Tabelle 7.1a und 7.1b angegebenen Vierfeldertafeln.

Tabelle 7.1a
Lösungswahrscheinlichkeit
für Person 1

Item j \ Item i	+	-	
+	.03	.27	.3
-	.07	.63	.7
	.1	.9	

Tabelle 7.1b
Lösungswahrscheinlichkeit
für Person 2

Item j \ Item i	+	-	
+	.63	.27	.9
-	.07	.03	.1
	.7	.3	

Tabelle 7.lc
Lösungswahrscheinlichkeit
für eine gemischte
Personenstichprobe

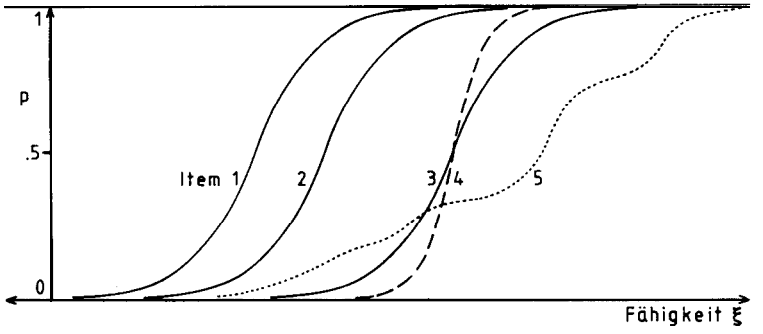
Item j \ Item i	+	-	
+	.33	.27	.6
-	.07	.33	.4
	.40	.60	

In jeder der beiden Vierfeldertafeln ist die Korrelation der beiden Items Null. Denkt man sich jedoch eine gemischte Personenstichprobe, bei der die Hälfte der Personen die Fähigkeitsausprägung ξ_1 hat, die andere Hälfte ξ_2, so ergibt sich für die Personengruppe die in Tabelle 7.lc angegebene Vierfeldertafel, die aus Tabelle 7.1a und 7.1b gemittelt ist. In Tabelle 7.lc korrelieren die Items i und j offensichtlich, und zwar zu $\rho=0.375$.

duzierbar" sind) vorgeschlagen. Näheres findet man bei Borg & Staufenbiel (1989, Kapitel 7).

In probabilistischen Latent-Trait-Modellen wird die deterministische Annahme, wonach nur Lösungswahrscheinlichkeiten von Null oder Eins vorkommen, durch eine probabilistische Annahme über die Itemcharakteristik ersetzt. Jedem Wert auf dem latenten Kontinuum ξ wird eine Wahrscheinlichkeit zugeordnet, mit der eine Person mit dieser Merkmalsausprägung das Item löst. Diese Funktion, die jedem Wert von ξ eine Lösungswahrscheinlichkeit zuordnet, heißt ***Itemcharakteristik*** des Items i und wird mit $p(i+/\xi)$ bezeichnet. Abbildung 7.2 zeigt Beispiele, wie Itemcharakteristiken aussehen können:

Abbildung 7.2: Itemcharakteristiken



Die Itemcharakteristiken der Items 1, 2, 3 entsprechen dem Rasch-Modell. Die Hinzunahme von Item 4 wäre im Birnbaum-Modell möglich. Item 5 hat eine unregelmäßig monoton steigende Itemcharakteristik.

Verschiedene Latent-Trait-Modelle unterscheiden sich darin, welche Form der Itemcharakteristik sie zulassen. In Abbildung 7.2 haben die Itemcharakteristiken der Items 1, 2 und 3 dieselbe Form und sind nur um einen bestimmten Betrag nach rechts oder links verschoben, was einer unterschiedlichen Itemschwierigkeit entspricht. Diese drei Items genügen dem einfachen Rasch-Modell (siehe Kapitel 7.2). Die Itemcharakteristik von Item 4 hat dieselbe allgemeine Form, jedoch einen steileren Anstieg, was einer größeren Trennscharfe entspricht. Items unterschiedlicher Trennschärfe sind im Birnbaum-Modell (siehe Kapitel 7.4) zulässig. Item 5 zeigt eine ebenfalls monoton steigende Itemcharakteristik, die aber keinem speziellen Latent-Trait-Modell entspricht. Eine weitere allen Latent-Trait-Modellen gemeinsame Annahme ist die ***lokale stochastische Unabhängigkeit*** der Items. Sie besagt, daß für jede einzelne Person (bei festem "Ort" auf dem latenten Kontinuum) die Beantwortung der Items stochastisch unabhängig erfolgt. Formal ausgedruckt: Die Wahrscheinlichkeit, bei gegebenem Personparameter von zwei Items i und j beide richtig zu lösen, ist das Produkt der Einzelwahrscheinlichkeiten:

$$[7.1] \qquad p(i+, j+/\xi) = p(i+/\xi)\, p(j+/\xi)$$

Lokale stochastische Unabhängigkeit besagt zunächst nichts darüber, wie in einer Gruppe von Personen mit beliebig verteilten Personparametern die Itemkorrelationen ausfallen. Sie werden im allgemeinen umso höher sein, je größer die Varianz der Per-

sonparameter ist. Beipiel 7.1 illustriert, wie bei lokaler stochastischer Unabhängigkeit die Itemkorrelationen aufgrund von Unterschieden in der latenten Dimension zustande kommen.

Im Unterschied zu unserem Beispiel werden in einer realen Population nicht nur zwei Werte von ξ vorkommen, sondern die Personparameter werden sich auf dem gesamten latenten Kontinuum verteilen. Je nachdem, wie diese Verteilung aussieht, ergibt sich höhere oder niedrigere Itemkorrelation. In einer Population ohne Varianz in den Personparametern ergibt sich eine Korrelation von Null. Anders ausgedruckt besagt also lokale Unabhängigkeit, daß alle Korrelationen zwischen den Items nur auf Unterschiede in der latenten Dimension zurückgehen dürfen. Weitere Abhängigkeiten (z.B. durch Faktoren, die nur bestimmten Itemgruppen gemeinsam sind) dürfen nicht bestehen. Da die lokale Unabhängigkeit somit beinhaltet, daß allen Items nur eine einzige gemeinsame latente Dimension zugrunde liegt, läßt sie sich als eine präzisere Fassung des Begriffs der Homogenität eines Tests verstehen.

Die Annahme eines latenten Kontinuums, der Begriff der Itemcharakteristik und die Annahme der lokalen stochastischen Unabhängigkeit sind allen Latent-Trait-Modellen gemeinsame Grundzüge. Die einzelnen Modelle unterscheiden sich in den Annahmen, die sie über die Form der Itemcharakteristik machen, und den daraus ableitbaren Folgerungen. Das im deutschen Sprachraum bekannteste Modell ist das einparametrige logistische Modell nach Rasch, das durch den von Fischer (1968) herausgegebenen Band “Testtheorie” bald Popularität gewann. Eine umfassende Darstellung des Rasch-Modells und der darauf gegründeten weiteren Entwicklung logistischer Modelle findet man bei Fischer (1974; 1983).

7.2 Das Rasch-Modell

Im Rasch-Modell ist jede Person v durch einen Personparameter 5_v und jedes Item i durch einen Itemparameter σ_i (griechisch: sigma. Der Itemparameter hat aber nichts mit dem Begriff der Standardabweichung zu tun) gekennzeichnet. Die Itemcharakteristik ist dann die logistische Funktion dieser beiden Parameter:

$$[7.2] \qquad p(i+/\xi_v,\sigma_i) = \frac{\exp(\xi_v - \sigma_i)}{1 + \exp(\xi_v - \sigma_i)}$$

Die Items 1, 2 und 3 in Abbildung 7.2 entsprechen dem Rasch-Modell. Die Form der Itemcharakteristik unterscheidet sich nur geringfügig von der Normalverteilungsfunktion, ist aber mathematisch leichter handhabbar. Ausgehend von einer Guttman-Skala kann man sich vorstellen, daß eine solche Itemcharakteristik zustande kommt, wenn die Ekken der Sprungfunktion durch Zufallseinflüsse (die Person oder auch das Item schwanken in ihrer Position auf dem Kontinuum) abgerundet werden.

Der Personparameter ist als der Ort der Person auf der latenten Dimension interpretierbar. Je größer der Personparameter ist (je weiter rechts die Person auf dem latenten Kontinuum plaziert ist), desto größer ist die Lösungswahrscheinlichkeit. Der Itemparameter drückt die Schwierigkeit eines Items aus: Die Itemcharakteristiken haben alle dieselbe Form, die nur nach rechts (schwierige Items) oder links (leichte Items) verschoben ist. Wenn man in Formel [7.2] $\sigma_i = 5_v$ einsetzt, so ergibt sich eine Lösungswahrscheinlichkeit von 0.5. Der Itemparameter σ_i gibt somit die Stelle des

latenten Kontinuums an, an der die Lösungswahrscheinlichkeit 0.5 ist. In Abbildung 7.2 hat von den drei Rasch-Items (1, 2 und 3) Item 3 den größten, Item 1 den kleinsten Schwierigkeitsparameter. Der Einfachheit der Darstellung wegen wird die Item-Charakteristik [7.2] bisweilen auch in der folgenden Form (delogarithmierte Parameter) geschrieben:

$$[7.3] \qquad p(i+/\Theta_v \varepsilon_i) = \frac{\Theta_v \varepsilon_i}{1 + \Theta_v \varepsilon_i}$$

mit $\Theta_v = \exp(\xi_v)$ und $\varepsilon_i = \exp(-\sigma_i)$
(Θ = griechisch: theta, ε = griechisch: epsilon)

Das Rasch-Modell weist folgende besonderen Vorzüge auf: Die Existenz erschöpfender Statistiken, die spezifische Objektivität der Parameterschätzung und die Möglichkeit darauf aufbauender Modellkontrollen. Das soll im folgenden kurz erläutert werden:

Im Rasch-Modell ist die Trefferzahl eine erschöpfende Statistik für den Personparameter. Praktisch bedeutet das, daß die gesamte Information über den Fähigkeitsgrad der Person in der Trefferzahl enthalten ist (eine formal präzise Darstellung des Begriffs findet man bei Fischer, 1974, Kapitel 12.2 - 12.4). Eine nähere Analyse des Antwortpatterns, um festzustellen, bei welchen Items die Person ihre Treffer erzielt hat, kann zu keiner verbesserten Schätzung ihres Fähigkeitsparameters führen, erübrigt sich also insoweit. Wenn ein Test dem Rasch-Modell entspricht, so ist damit die häufig nur der Einfachheit wegen gewählte Auswertungsart, wonach die Anzahl der Richtigen festgestellt und als Testrohwert verwendet wird, auch als die optimale Auswertung theoretisch begründet.

Der mathematische Beweis dafür, daß die Trefferzahl tatsächlich die gesamte Information enthält, die man über den Personparameter gewinnen kann, kann hier nur skizziert werden. Er wird geführt, indem man zeigt, daß bei gegebener Trefferzahl die bedingte Wahrscheinlichkeit für die einzelnen Antwortpatterns nicht vom Personparameter abhängt, sondern nur von den Itemparametern - anders gesagt: keine Information über den Personparameter, sondern nur über die Itemparameter enthält. Für den Fall, daß der Test nur aus zwei Items besteht, ist dieser Beweis leicht zu führen. In Beispiel 7.2 wird gezeigt, daß die Wahrscheinlichkeit, daß eine Person mit 1 Treffer das erste (und nicht das zweite) Item gelöst hat, nicht vom Personparameter abhängt, sondern nur von den beiden Itemparametern. Das Ergebnis läßt sich auf mehr als zwei Items verallgemeinem: die bedingte Wahrscheinlichkeit, daß bei r Treffern die einzelnen Items gelöst/nicht gelöst sind, hängt nur vom Verhältnis der Itemschwierigkeiten, nicht aber vom Personparameter ab. Den vollständigen Beweis mit beliebig vielen Items und Personen findet man bei Fischer, 1974, Kapitel 13.5.

Parameterschätzung und Modellkontrollen: Die zwar rechnerisch aufwendigste, aber theoretisch am besten begründete Methode zur Schätzung der Itemparameter ist die CML-Schätzung (Conditional-Maximum-Likelihood-Schätzung). Dabei wird die Eigenschaft des Modells, daß die Information über die Personparameter (enthalten in der Trefferzahl) von der Information über die Itemparameter (enthalten in der Verteilung der Treffer auf die Items bei gegebener Trefferzahl) separierbar ist, voll genutzt. Es wird nur die von den Personparametern unabhängige Information zur Schätzung der Itemparameter verwendet (die mathematische Ableitung und die rechnerische Durchführung der CML-Schätzung sind aufwendig und können hier nicht dargestellt

Beispiel 7.2: Berechnung der bedingten Wahrscheinlichkeit, daß eine Person v, die bei zwei Items r = 1 Treffer erzielt hat, diesen Treffer beim ersten (und nicht beim zweiten) Item erzielt.

Wir gehen von der Itemcharakteristik in der Schreibweise von Formel [7.3] aus und berechnen zunächst die Wahrscheinlichkeit, daß die Person v das Item 1 löst und das Item 2 verfehlt, also das Antwortmuster "10" liefert. Aufgrund der lokalen Unabhängigkeit ist das das Produkt der beiden Einzelwahrscheinlichkeiten:

$$p(10/\Theta_v) = \frac{\Theta_v \varepsilon_1}{1+\Theta_v \varepsilon_1} \cdot \left(1 - \frac{\Theta_v \varepsilon_2}{1+\Theta_v \varepsilon_2}\right) = \frac{\Theta_v \varepsilon_1}{(1+\Theta_v \varepsilon_1)(1+\Theta_v \varepsilon_2)}$$

Des weiteren brauchen wir die Wahrscheinlichkeit, daß die Person v genau einen Treffer erzielt. Dazu haben wir die Wahrscheinlichkeiten für die beiden Möglichkeiten, die zu r = 1 Treffer führen, nämlich das Antwortmuster "10" und "01", zu addieren.
Für die Wahrscheinlichkeit, daß die Person das Antwortmuster "01" erzielt, erhalten wir (Ableitung analog zur Rechnung für "10"):

$$p(01/\Theta_v) = \frac{1}{1+\Theta_v \varepsilon_1} \cdot \frac{\Theta_v \varepsilon_2}{1+\Theta_v \varepsilon_2}$$

Die Addition der beiden Möglichkeiten ergibt:

$$p(r=1/\Theta_v) = p(10/\Theta_v) + p(01/\Theta_v)$$

Die bedingte Wahrscheinlichkeit, daß die Person v das Antwortmuster "10" hat, wenn sie r = 1 Treffer erzielt hat, erhält man, indem man die entsprechenden Wahrscheinlichkeiten dividiert (Anteil der Fälle mit Muster "10" an der Gesamtheit aller Fälle, die zu r = 1 führen):

$$p(10/r=1,\Theta_v) = \frac{p(10/\Theta_v)}{p(r=1/\Theta_v)}$$

$$= \frac{\Theta_v \varepsilon_1}{(1+\Theta_v \varepsilon_1)(1+\Theta_v \varepsilon_2)} \cdot \frac{(1+\Theta_v \varepsilon_1)(1+\Theta_v \varepsilon_2)}{\Theta_v (\varepsilon_1+\varepsilon_2)}$$

Nach Kürzen erhält man dann das Ergebnis:

$$p(10/r=1,\Theta_v) = \frac{\varepsilon_1}{\varepsilon_1+\varepsilon_2}$$

Man sieht, daß alle Ausdrucke, die den Personparameter enthalten, durch Kürzen weggefallen sind. Die bedingte Wahrscheinlichkeit, daß Item 1 richtig ist, wenn insgesamt r = 1 Treffer erzielt wurde, hängt somit nicht vom Personparameter ab, sondern nur von den Itemparametern. Bei bekannter Trefferzahl (hier: r = 1) enthält somit das Antwortmuster (Item 1, nicht Item 2 wurde gelöst) keine weitere (d.h. über die Trefferzahl hinausgehende) Information über den Personparameter, sondern lediglich Information über die Itemparameter.

werden. Die Durchführung erfordert auch bei kleineren Itemzahlen EDV-Einsatz. Näheres findet man bei Fischer, 1974, Kapitel 14). Als Folge davon hängt das Ergebnis einer CML-Schätzung der Itemparameter auch nicht davon ab, wie sich in der speziellen Stichprobe die Personparameter verteilen. Praktisch bedeutet das, daß eine CML-Schätzung (im Rahmen der Schätzgenauigkeit) immer zum selben Ergebnis führen muß, egal an welcher Teilstichprobe von Personen sie vorgenommen wird. Außerdem muß die Schätzung (wieder abgesehen von Fragen der Schätzgenauigkeit) immer zum selben Ergebnis führen, wenn sie für eine beliebige Teilmenge von Items vorgenommen wird. Diese Eigenschaft des Modells (daß die Itemparameter unabhängig von den Personparametern geschätzt werden können und daß sie sich nicht ändern, wenn modellkonforme Items hinzugefügt oder weggelassen werden) nennt man ***spezifische Objektivität*** (früher wurde bisweilen der irreführende Ausdruck "Populationsunabhängigkeit" verwendet).

Die Prüfung der Modellgeltung baut auf der spezifischen Objektivität der CML-Schätzung auf. Das Datenmaterial wird auf möglichst viele verschiedene Arten (z.B. nach der Trefferzahl in Personen mit überdurchschnittlicher versus unterdurchschnittlicher Trefferzahl; oder danach, ob sie ein bestimmtes Item gelöst/ nicht gelöst haben; oder nach verschiedenen Außenkriterien wie Alter, Geschlecht, Schulbildung usw.) unterteilt und jeweils aus den verschiedenen Teil-Datensätzen getrennt die Item-Parameter geschätzt. Mit Hilfe von Signifikanztests kann überprüft werden, ob die CML-Schätzungen voneinander verschieden sind, was bei Modellgeltung nicht der Fall sein darf. Sofern nur bei einzelnen Items Differenzen auftreten, kann man diese Items eliminieren und erneut prüfen, ob die verbleibenden Items nunmehr eine Raschhomogene Skala bilden. Diese Überprüfung sollte - wie immer, wenn eine Testrevision anhand der Daten erfolgt ist - an neuem, unabhängigen Datenmaterial erfolgen.

Neben der Methode der CML-Schätzung für die Itemparameter und den darauf aufbauenden Signifikanztests zur Modellkontrolle gibt es eine Reihe anderer Parameter-Schätzverfahren und andere Methoden zur Prüfung der Modellgeltung. Diese sind z.T. rechnerisch einfacher, aber theoretisch weniger gut begründet (Näheres siehe Fischer, 1974, 1983).

Im Unterschied zum Testautor interessieren den Testanwender weniger die Item-Parameter als die Personparameter. Wenn die Itemparameter bekannt sind, können die Personparameter aus den Trefferzahlen geschätzt werden. Da die Schätzwerte für die Personparameter letztlich nur eine monotone Transformation der Trefferzahl sind (je mehr Treffer, desto höher der geschätzte Personparameter), ist im allgemeinen wohl nicht zu erwarten, daß sich an den Korrelationen des Tests mit Außenkriterien viel ändert, wenn man die geschätzten Personparameter anstelle der Trefferzahl zur Vorhersage benutzt. Das zeigte sich z.B. beim Mannheimer Test zur Erfassung des physikalisch-technischen Problemlösens (MTP von Conrad, Baumann & Mohr, 1980), bei dem sowohl für die Trefferzahl als auch für die geschätzten Personparameter Kriteriumskorrelationen berechnet wurden. Die Unterschiede in den Korrelationen waren gering und unsystematisch.

Vom Modellansatz her eignet sich das einfache Rasch-Modell besonders für Leistungstests ohne wesentliche Speed-Komponente und für Fragebogen mit nur zwei Antwortmöglichkeiten. Beispiele für Anwendungen in unterschiedlichen Bereichen sind bei Fischer (1974, 1983) referiert. Publizierte Tests, bei denen neben der Analyse nach der klassischen Testtheorie auch Rasch-Analysen der Items durchgeführt wurden und Umrechnungstabellen von Rohwerten in geschätzte Personparameter

angegeben sind, sind u.a. der oben genannte Mannheimer Test zur Erfassung des physikalisch-technischen Problemlösens von Conrad et al.(1980), der Anstrengungsvermeidungstest von Rollett & Bartram (1977) und das Adaptive Intelligenz-Diagnostikum von Kubinger & Wurst (1980).

7.3 Das linear-logistische Modell

Im einfachen Rasch-Modell ist jedes Item durch einen Parameter gekennzeichnet, der die Schwierigkeit des Items angibt. Im linear-logistischen Modell wird dieser Parameter in additive Anteile zerlegt, die für das Zustandekommen der Itemschwierigkeit verantwortlich sind. So kann die Lösung einer Aufgabe mehrere Teilschritte (z.B. Anwendung bestimmter Regeln) erfordern, von denen jeder zur Schwierigkeit beiträgt. Ziel ist es, nicht nur die Aufgabenschwierigkeiten anzugeben, sondern auch die Schwierigkeiten der einzelnen für die Lösung erforderlichen Operationen.

Wie beim einfachen Rasch-Modell ist die Wahrscheinlichkeit, daß Proband v Item i löst, durch die logistische Funktion beschrieben:

$$p(i + /\xi_v, \sigma_i) = \frac{\exp(\xi_v - \sigma_i)}{1 + \exp(\xi_v - \sigma_i)}$$

Der Itemparameter σ_i seinerseits wird als lineare Funktion der sogenannten "Basisparameter" η_j erklärt:

[7.4] $$\sigma_i = \sum_j q_{ij}\, \eta_j + c$$

mit σ_i = Itemschwierigkeitsparameter für Item i
η_j = Schwierigkeit von Operation j
q_{ij} = Gewichtszahl für das Vorkommen von Operation j in Item i (s. unten)
c = beliebig wählbare Normierungskonstante (z.B. die Zahl Null). Ihre Hinzunahme in Formel [7.4] drückt aus, daß die Itemparameter nur bis auf eine additive Konstante bestimmt sind, also auf einer Differenzenskala liegen.

So z.B. analysierte Spada (1976) die Schwierigkeit von Aufgaben aus dem Bereich der Mechanik, u.a. Aufgaben zur Übertragung von Drehrichtungen in Räderwerken. Zur Lösung der Aufgaben waren 6 Regeln anzuwenden, z.B.:

Regel 1: Zwei mit ihren Umfängen aufeinander stoßende Räder haben gegenläufige Drehrichtung.
Regel 2: Zwei Räder, die fest auf einer gemeinsamen Achse sitzen, haben gleichläufige Drehrichtung.
... usw.

Jede Aufgabe zeigte ein mehr oder weniger komplexes Räderwerk, so daß zur Lösung mehrere Regeln erforderlich sein und auch einzelne Regeln mehrfach zur Anwendung kommen konnten. In Formel [7.4] sind dann die η_j die Schwierigkeiten

der einzelnen Regeln und q_{ij} die Häufigkeiten, mit denen die Regeln in den einzelnen Items angewandt werden müssen.

Die Basisparameter des linear-logistischen Modells lassen sich ebenfalls mit der CML-Methode spezifisch objektiv schätzen. Die Hypothesen über die Zusammensetzung der Itemschwierigkeiten aus den Basisparametern können geprüft werden, indem man die Itemschwierigkeiten aus dem einfachen Rasch-Modell schätzt und damit die Schwierigkeiten vergleicht, die sich aufgrund der Schwierigkeitsbeiträge der beteiligten Operationen ergeben müßten. Die Signifikanz der Abweichungen kann mit einem Likelihood-Quotienten-Test geprüft werden. Weitere Prüfungen sind möglich, indem man die Basisparameter aus verschiedenen Teilmengen von Personen und Items schätzt (Näheres siehe Scheiblechner, 1975; Nährer, l980). Aufgrund der spezifischen Objektivität müßten die Schätzungen im Rahmen der Schätzgenauigkeit übereinstimmen.

Anwendungen des linear-logistischen Modells: Das linear-logistische Modell bietet sich an, wenn es darum geht, einen Stoff abzufragen, der die Anwendung einer begrenzten Zahl von Regeln erfordert. Eine Reihe von Anwendungen dieser Art ist bei Fischer (1974) dargestellt. Im Idealfall könnte es gelingen, die Determinanten der Itemschwierigkeiten voll aufzuklaren und einen beliebig großen Itempool mit Items bekannter Schwierigkeit zu generieren. Dieses Ziel ist allerdings bisher nirgends voll erreicht worden:

Spada (1976) mußte einige seiner Aufgaben zur Mechanik nachträglich ausscheiden, weil sie dem Rasch-Modell nicht genügten. Die Erklärung der Itemschwierigkeiten aus den Basisparametern gelang nur unvollständig: Die Abweichungen der Itemschwierigkeiten von den aufgrund der Basisparameter vorhergesagten Werten waren zwar numerisch gering, aber signifikant. Die Schätzung der Basisparameter aus verschiedenen Teilstichproben ergaben ebenfalls zum Teil signifikante Unterschiede.

Verschiedene Autoren befaßten sich mit der Analyse von Items nach dem Vorbild des Matrizen-Tests: So konstruierte Formann (1973, zitiert nach Hornke & Habon, 1984) 42 Items, deren Schwierigkeiten er auf 10 Basisparameter (Art der zu erkennenden Regel, Richtung der Regelgeltung, verschiedene Materialeigenschaften usw.) zurückführte. Nährer (1980) versuchte für 10 neu konstruierte Items die Schwierigkeiten aus Formanns Basisparametern vorherzusagen. Aufgrund seiner Daten schlug er eine modifizierte Operationsstruktur vor, die auch für Formanns Daten eine verbesserte Anpassung brachte. Hornke & Habon (1984) versuchten Nährers (1980) Ergebnisse zu replizieren, indem sie 12 von Nährers Items zusammen mit 14 neu konstruierten analysierten. Dabei ergaben sich schon beim einfachen Rasch-Modell z.T. signifikante Abweichungen. Die Schätzungen für die Basisparameter stimmten mit den Angaben Nährers nur zum Teil überein, die Itemschwierigkeiten wichen von den aus den Basisparameter berechneten Werten signifikant ab. Erst bei nachträglichem Ausschluß von 6 Items verbesserte sich das Bild.

Ähnlich erging es Gittler (1984) bei dem Versuch, Würfelaufgaben zur Erfassung des räumlichen Vorstellens (ähnlich dem Subtest “Würfel” im Intelligenz-Struktur-Test von Amthauer, 1970) in ihrer Schwierigkeit zu erklären. Als Ergebnis eines längeren formalen und inhaltlichen Analyseprozesses gelang es ihm schließlich, 17 Items zu finden, die dem Rasch-Modell genügten und deren Schwierigkeiten er auf 9 Basisparameter (Zahl der erforderlichen Lösungsschritte, Musterkombination, Plazierung der Lösung unter den Antwortalternativen usw.) zurückführen konnte. In einer Kreuzvalidierung (Überprüfung an neuen, unabhängigen Daten) waren die 17 Items

wieder Rasch-konform, und es ließ sich wieder dieselbe linear-logistische Modellstruktur mit 9 Basisparametern anpassen. Die Schätzwerte für die Parameter stimmten aber nur zum Teil überein, was Gittler auf den unterschiedlichen Aufgabenkontext (die 17 Aufgaben waren zwischen anderen eingebettet) zurückführt.

Zusammenfassend läßt sich feststellen, daß es zwar immer wieder gelingt, einen Rasch-konformen Itemsatz zu finden und die Itemschwierigkeiten auf Basisparameter zurückzuführen. Die Versuche, die Basisparameter an anderen Stichproben numerisch zu replizieren oder die Schwierigkeiten neuer Items vorherzusagen, haben allerdings nur begrenzte Erfolge gehabt. Das Ziel, in ihrer Schwierigkeit perfekt vorhersagbare Itempools zu konstruieren, steht noch in weiter Feme, wenn es überhaupt realistisch ist.

7.4 Das mehrkategoriale Rasch-Modell

Der Ansatz des Rasch-Modells läßt sich auf Items mit mehr als zwei Antwortkategorien verallgemeinem. So könnte man z.B. einen Interessenstest für 4 Interessensrichtungen (z.B. Kunst, Naturwissenschaften usw.) konstruieren, bei dem den Probanden in jedem Item 4 Tätigkeiten zur Wahl vorgelegt werden. Jede der 4 Tätigkeiten entstammt einem anderen der vier Interessensgebiete, und bei der Auswertung des Tests wird ausgezählt, wie oft sich der Proband für jedes Gebiet entschieden hat. Da die Gesamtzahl der Wahlen der Itemzahl entsprechen muß, kann auf diese Art nur die relative Ausprägung der Interessen untereinander zum Ausdruck kommen: Kein Proband kann auf allen Interessensrichtungen hohe oder auf allen Interessensrichtungen niedrige Werte haben, auch wenn er sich für alle vier Gebiete sehr stark oder für alle vier Gebiete sehr wenig interessiert (das Beispiel ist an den Berufs-lnteressen-Test BIT von Irle und Allehoff, 1984, angelehnt. Die Art der Itemkonstruktion beim BIT ist aber komplizierter, da 9 Interessensrichtungen mit Hilfe von Items mit 4 Wahlalternativen abgefragt werden). Jede Person v ist dann durch 4 Personparameter ($\xi_v^{(1)}, \xi_v^{(2)}, \ldots \xi_v^{(4)}$) gekennzeichnet, die ihre Tendenz ausdrucken, sich für jedes der 4 Interessensgebiete zu entscheiden. Analog dazu ist jedes Item durch 4 Itemparameter ($\sigma_i^{(1)}, \sigma_i^{(2)}, \ldots \sigma_i^{(4)}$) gekennzeichnet, die die "Schwierigkeit" (Unattraktivität) der Alternativen (Interessensgebiete) in diesem Item ausdrücken. Die Wahrscheinlichkeit, daß die Person v bei Item i das Interessensgebiet g wählt, soll sich gemäß den Modellannahmen wie folgt ergeben:

$$[7.5] \qquad p(g+/v, i) = \frac{\exp(\xi_v^{(g)} - \sigma_i^{(g)})}{\sum_j \exp(\xi_v^{(j)} - \sigma_i^{(j)})}$$

Bei Formel [7.5] wurde die in der Rasch-Literatur übliche Notation übernommen. Die in Klammem hochgestellten Indizes sind keine Exponenten, sondern werden lediglich hochgestellt, um im Fußraum mehr Platz zu behalten. Um sie von Exponenten zu unterscheiden, sind sie eingeklammert.

Ähnlich wie beim einfachen, zweikategorialen Rasch-Modell gibt es auch im mehrkategorialen Rasch-Modell erschöpfende Statistiken: Die Häufigkeiten, mit denen sich eine Person für die einzelnen Interessensrichtungen entschieden hat, sind erschöpfende Statistiken für ihre Personparameter. Wenn sich ein Proband bei 20 Items 9 mal für das Interessensgebiet "Kunst" entschieden hat, so ist in dieser "Trefferzahl"

die gesamte Information über seine Interessensausprägung (relativ zu den Interessen in den anderen Gebieten) enthalten. Es erübrigt sich, näher zu analysieren, bei welchen Items er "Kunst" gewählt/nicht gewählt hat.

Die Häufigkeiten, mit denen die vier Alternativen eines Items gewählt wurden, sind erschöpfende Statistiken für die Itemparameter. Ähnlich wie beim zweikategorialen Rasch-Modell stehen auch beim mehrkategorialen Rasch-Modell zur Schätzung der Itemparameter CML-Schätzverfahren zur Verfügung, die eine spezifisch objektive Schätzung (siehe Kapitel 7.3) der Itemparameter ermöglichen. Auch hier kann die Modellgeltung geprüft werden, indem man den Datensatz nach unterschiedlichen Gesichtspunkten (Personen mit hohen/ niedrigen Punktwerten in der Interessensrichtung "Kunst"; nach Außenkriterien wie Geschlecht, Alter, Schulnoten usw.) unterteilt und in den Teilstichproben getrennt die Itemparameter schätzt. Bei Modellgeltung müssen die aus den verschiedenen Datensätzen gewonnenen Schätzungen für die Itemparameter (im Rahmen der Schätzgenauigkeit) übereinstimmen. Letzteres kann mit Hilfe von Signifikanztests geprüft werden.

Zur Interpretation der Parameter: Aufgrund der Aufgabenstellung, bei der die Person genau eine der vier Interessensrichtungen zu wählen hat, ist offensichtlich, daß das Testergebnis nicht eine Angabe über die absolute Höhe der Interessensausprägungen in den einzelnen Gebieten sein kann, sondern nur eine Angabe über das relative Überwiegen der einzelnen Interessensrichtungen gegenüber den anderen. Bei einer Person, die in allen Gebieten hohe Interessen hat, können sich die Wahlen genauso verteilen wie bei einer anderen, die an allen Gebieten wenig Interesse hat.

Die Tatsache, daß die Daten keine Information über die absolute Höhe der Interessensausprägung enthalten, sondern nur über die relative Höhe der Interessensausprägung in einem Gebiet gegenüber den anderen Gebieten, druckt sich im Modell darin aus, daß die vier Personparameter nur bis auf eine frei wählbare additive Konstante bestimmt sind. Man kann diese Konstante z.B. so wählen, daß der Mittelwert der vier Personparameter für jede Person Null ist. Die Stärke jeder Interessensrichtung wird dann relativ zum Durchschnitt aller vier Interessen angegeben.

Analoges gilt für die Itemparameter: Aus den Daten erfährt man, wieviele Personen sich für die einzelnen Alternativen entschieden haben. Daraus ist aber nicht ersichtlich, ob alle vier Alternativen hoch attraktiv oder unattraktiv waren, sondern nur die relative Attraktivität der einzelnen Alternative im Vergleich zu den anderen. Dementsprechend sind auch die Itemparameter nur bis auf eine additive Konstante festgelegt. Auch hier erscheint es naheliegend, für jedes Item den Mittelwert der Itemparameter auf Null festzulegen und damit die Attraktivität jeder Alternative relativ zur durchschnittlichen Attraktivität aller vier Alternativen anzugeben.

Beispiel 7.3 illustriert an einem Zahlenbeispiel den durch Formel [7.5] ausgedrückten Zusammenhang zwischen den Parametern und den Wahlwahrscheinlichkeiten für die einzelnen Alternativen und die beliebige Wahl einer Normierungskonstanten für die Person- und Itemparameter.

Im vorliegenden Beispiel der vier Interessensrichtungen sind die vier Antwortkategorien offensichtlich qualitativ verschieden. In anderen Fällen kann sich die Frage stellen, ob sich die Kategorien nicht ordnen und auf nur eine Dimension zurückführen lassen: So könnten z.B. in einem Fragebogen die Antwortmöglichkeiten "Ja/ ? / Nein" Ausdruck unterschiedlich starker Zustimmung sein, oder es könnten sich zunächst für qualitativ gehaltene Kategorien (z.B. intropunitive, impunitive und extrapunitive Reaktionen im Rosenzweig Picture-Frustration-Test nach Rauchfleisch,

Beispiel 7.3: Berechnung der Wahlwahrscheinlichkeiten für die einzelnen Antwortalternativen eines Items im mehrkategorialen Rasch-Modell

Wir nehmen an, eine Person v habe für vier Interessensrichtungen folgende Personparamet

$$\xi_v^{(1)} = 0,\ \xi_v^{(2)} = +1,\ \xi_v^{(3)} = -2,\ \xi_v^{(4)} = +3$$

Die Itemparameter für Item i ("Schwierigkeit" oder Unattraktivität der für die einzelnen Interessensgebiete angebotenen Alternativen, von denen die Person eine ankreuzen muß) seien:

$$\sigma_i^{(1)} = +1,\ \sigma_i^{(2)} = +2,\ \sigma_i^{(3)} = 0,\ \sigma_i^{(4)} = +3$$

(a) Man berechne nach Formel [7.5] die Wahrscheinlichkeiten, mit der sich Person v bei Item i für die einzelnen Alternativen entscheidet.
(b) Man normiere Personparameter und Itemparameter jeweils auf den Mittelwert Null und führe die Berechnung nach Formel [7.5] erneut durch.
Lösung:
a) Man berechne zunächst für jede Kategorie j den Ausdruck $\exp((\xi_v^{(j)} - \sigma_i^{(j)})$:

Kategorie 1: $\exp(0 - 1) = 0.3679$
" 2: $\exp(1 - 2) = 0.3679$
" 3: $\exp(-2 - 0) = 0.1353$
" 4: $\exp(3 - 3) = 1.0000$
$\Sigma \exp(\xi_v^{(j)} - \sigma_i^{(j)}) = 1.8708$

Damit erhält man gemäß Formel [7.5] die Wahlwahrscheinlichkeiten für die Kategorien:

Kategorie 1: 0.3679/1.8708 = 0.1966
" 2: 0.3679/1.8708 = 0.1966
" 3: 0.1353/1.8708 = 0.0723
" 4: 1.0000/1.8708 = 0.5345

Man sieht, die Wahlwahrscheinlichkeit ist für Kategorie 4 am größten, weil hier der Personparameter relativ zum Itemparameter am größten ist (die Differenz Personparameter minus Itemparameter ist bei den Kategorien 1 bis 3 negativ, bei Kategorie 4 Null).

b) Um beide Parametergruppen jeweils auf den Mittelwert Null zu normieren, ziehen wir von den Personparametern die Zahl 0.5, von den Itemparametern 1.5 ab. Die Werte für die Parameter lauten dann:
$\xi_v^{(1)} = -.5,\ \xi_v^{(2)} = +.5,\ \xi_v^{(3)} = -2.5,\ \xi_v^{(4)} = +2.5$
$\sigma_i^{(1)} = -.5,\ \sigma_i^{(2)} = +.5,\ \sigma_i^{(3)} = -1.5,\ \sigma_i^{(4)} = +1.5$
Als nächstes berechnen wir wieder für jede Kategorie den Ausdruck $\exp(\xi_v^{(j)} - \sigma_i^{(j)})$:

Kategorie 1: $\exp(-.5 - (-.5)) = 1.000$
" 2: $\exp(+.5 - (+.5)) = 1.000$
" 3: $\exp(-2.5 - (-1.5)) = 0.368$
" 4: $\exp(+2.5-(+1.5)) = 2.718$
$\Sigma \exp(\xi_v^{(j)} - \sigma_i^{(j)}) = 5.086$

Daraus ergeben sich die Wahlwahrscheinlichkeiten für die Kategorien als:

Kategorie 1: 1/5.086 = 0.1966
" 2: 1/5.086 = 0.1966
" 3: .368/5.086 = 0.0723
" 4: 2.718/5.086 = 0.5344

Die Wahlwahrscheinlichkeiten sind also gegenüber der ersten Berechnung unverändert. Das Hinzufügen einer Konstanten (hier: des Mittelwerts) zu allen Personparametern einer Person oder zu allen Itemparametern eines Items ändert nichts an den Wahlwahrscheinlichkeiten. Anders gesagt: Die Personparameter (analog: Itemparameter) sind durch Formel [7.5] nur bis auf eine beliebig wählbare additive Konstante bestimmt.

1979) als Abstufungen nur einer Dimension erweisen. Solche Hypothesen können in Anschluß an die Prüfung der Modellgeltung für das mehrkategoriale Rasch-Modell als speziellere Hypothesen über die Parameter ausgedrückt und getestet werden.

7.5 Das Birnbaum-Modell

Während das einfache Rasch-Modell nur einen Itemparameter enthält, der die Schwierigkeit des Items ausdrückt und die Itemcharakteristik nach rechts oder links verschiebt (siehe Abbildung 7.1), enthält das Birnbaum-Modell einen zweiten Item-Parameter, der die Itemcharakteristiken bei sonst gleicher Form steiler oder flacher ansteigen läßt. Ein steilerer Anstieg entspricht einer größeren Trennschärfe des Items, weshalb dieser Parameter auch als Trennschärfeparameter bezeichnet wird. In Abbildung 7.2 bilden die Items 1, 2, 3 und 4 eine Birnbaum-Skala, wobei die Items 1, 2 und 3 einen flacheren Anstieg der Itemcharakteristik zeigen als Item 4, bei dem die Itemcharakteristik einen steileren Verlauf zeigt. Die Items 1, 2 und 3 haben denselben Trennschärfeparameter, Item 4 hat einen größeren. Im Birnbaum-Modell ist nicht die Summe der richtigen Lösungen die erschöpfende Statistik für die Personparameter, sondern es ist eine gewichtete Summe zu bilden, wobei die Gewichtszahlen den Trennschärfeparametern der Items entsprechen, so daß trennscharfe Items höher gewichtet werden als weniger trennscharfe.

Darüber hinaus wurden verschiedene Versuche gemacht, auch Ratewahrscheinlichkeiten mit einzubeziehen und Strategien zur Behandlung ausgelassener Antworten entwickelt. Die hierfür verfügbaren Rechenprogramme wurden überwiegend in den USA entwickelt und berücksichtigen mehr pragmatische als theoretische Gesichtspunkte. Einen Überblick über die Schätzverfahren und eine vergleichende Diskussion von zwei Rechenprogrammen findet man bei Swaminathan & Gifford (1983). Weiterhin liegen Erfahrungsberichte zur Stabilität der Schätzungen auch bei nicht modellkonformen Daten vor. So berichten Goldman & Raju (1986) über eine Studie an realen und an simulierten Daten, in der die Schätzwerte für die Personparameter nahezu perfekt korrelierten, wenn sie der Auswertung einmal das einfache Rasch-Modell, das andere Mal das zweiparametrige Birnbaum-Modell zugrunde legten. Zum gleichen Ergebnis kamen Hambleton & Cook (1983), die Simulationsstudien mit dem ein-, zwei- und dreiparametrigen Modell machten. Die Schätzung der Personparameter verschlechterte sich kaum, wenn der Analyse das einfache Rasch-Modell zugrundegelegt wurde, obwohl das zwei- oder dreiparametrige Modell zutraf.

Übersicht 7.1: Die wichtigsten Varianten logistischer Modelle

Rasch-Modell (wird auch "einparametriges logistisches Modell" genannt)
- Antwortmöglichkeiten: 2 (richtig/falsch)
- Personparameter: 1 (Fähigkeit)
- Itemparameter: 1 (Schwierigkeit)

Linear-logistisches Modell
- Antwortmöglichkeiten: 2 (richtig/falsch)
- Personparameter: 1 (Fähigkeit)

Itemparameter: 1 (Schwierigkeit) Dieser Parameter wird als gewichtete Summe von Basisparametern (Schwierigkeit von beteiligten Operationen) erklärt

Mehrkategoriales Rasch-Modell
- Antwortmöglichkeiten: k (eine von k Kategorien ist anzukreuzen)
- Personparameter: k (Tendenz der Person eine bestimmte Kategorie zu wählen; relative Bevorzugungstendenz gegenüber den anderen Kategorien)
- Itemparameter: k ("Schwierigkeiten" der Kategorien bei diesem Item, relatives Ausmaß in dem das Item eine jede Reaktionskategorie provoziert)

Birnbaum-Modell (wird auch "zweiparametriges logistisches Modell" genannt)
- Antwortkategorien: 2 (richtig/falsch)
- Personparameter: 1 (Fähigkeit)
- Itemparameter: 2 (Schwierigkeit, Trennscharfe)

Dreiparametriges logistisches Modell
- Antwortkategorien: 2 (richtig/falsch)
- Personparameter: 1 (Fähigkeit)
- Itemparameter: 3 (Schwierigkeit, Trennscharfe, Rateparameter)

7.6 Dem Latent-Trait-Ansatz verwandte Modelle

Das linear logistische Modell mit gelockerten Annahmen (LLRA-Modell = Linear Logistic Model with Relaxed Assumptions)

Dieses Modell setzt voraus, daß für jedes Item die Itemcharakteristik die im Rasch-Modell angenommene Form hat. Es macht jedoch keinerlei Annahmen über die Dimensionalität: Jedes Item kann von einer anderen latenten Dimension abhängen, und der Proband kann durch ebensoviele Personparameter gekennzeichnet sein, wie Items vorhanden sind. Ziel ist es, in Vorher-Nachher-Versuchsplänen Behandlungseffekte zu schätzen. Da es hier nicht darum geht, Personen Meßwerte zuzuordnen, ist das Modell auch nicht zur Testtheorie zu rechnen.

Das Latent-Class-Modell

Der theoretische Ansatz des Latent-Class-Modells ist dem der Latent-Trait-Modelle in vielerlei Hinsicht verwandt, wobei an die Stelle der quantitativen latenten Dimension eine Einteilung der Personen in qualitativ verschiedene Klassen tritt. Diese Klassen sind nicht direkt beobachtbar (latent). Die Wahrscheinlichkeit, daß eine Person ein Item löst, hängt davon ab, in welche Klasse die Person gehört. Innerhalb jeder Klasse sind die Items unabhängig (lokale Unabhängigkeit), und alle Abhängigkeiten, die man zwischen den Items findet, gehen darauf zurück, daß die Personenstichprobe aus unterschiedlichen Klassen zusammengesetzt ist. Ziel der Analyse ist es herauszufinden, wieviele latente Klassen es gibt, und die Lösungswahrscheinlichkeiten für die einzelnen Items anzugeben. Für den einzelnen Probanden kann dann anhand seines Antwortmusters berechnet werden, mit welcher Wahrscheinlichkeit er den ein-

zelnen Klassen zuzurechnen ist. Dieser Modellansatz soll im folgenden anhand einer Arbeit von Formann, Ehlers & Scheiblechner (1980) illustriert werden, die hier allerdings nur in Auszügen widergegeben werden kann.

Formann et al. (1980) wendeten die Latent-Class-Analyse auf die Daten der Eichstichprobe zur Marburger Verhaltensliste (MVL von Ehlers, Ehlers & Makus, 1978) an. Die MVL enthält fünf Skalen zur Diagnose verhaltensauffälliger Kinder. Neben verschiedenen Latent-Class-Analysen zu den einzelnen Skalen (über die hier nicht berichtet wird) wurden auch mehrere Latent-Class-Analysen mit Items aus verschiedenen Unterskalen durchgeführt. Das Ergebnis einer dieser Analysen wird im folgenden etwas vereinfacht dargestellt:

Aus den fünf Unterskalen der MVL wurden drei Unterskalen, nämlich "Instabiles Leistungsverhalten (IL)", "Unangemessenes Sozialverhalten (US)" und "Kontaktangst (KA)" herausgegriffen, und aus jeder dieser Skalen zwei besonders gute Items ausgewählt, insgesamt also 6 Items. Für jedes Item wurden zwei Antwortkategorien gebildet (Symptom wurde beobachtet: ja/nein). Bei 6 Items gibt es dann 2^6= 64 mögliche Antwortmuster. Deren Häufigkeiten in der Eichstichprobe von n = 1172 Schülern wurden ausgezählt und bildeten die Datenbasis für die Latent-Class-Analyse.

Als Ergebnis der Latent-Class-Analyse fand man, daß sich die 64 Häufigkeiten erklären lassen, wenn man annimmt, daß es vier latente Klassen gibt, in denen die einzelnen Symptome die in Tabelle 7.1 angegebenen Auftretenswahrscheinlichkeiten haben.

Tabelle 7.1: Ergebnis einer Latent-Class-Analyse von sechs Items aus der Marburger Verhaltensliste (nach Formarm et al., 1980). Auftretenswahrscheinlichkeiten der Symptome in den vier latenten Klassen und relative Anteile der Klassen in der Eichstichprobe der Marburger Verhaltensliste.

	Klasse 1	Klasse 2	Klasse 3	Klasse 4
Item				
IL 1	.87	.57	.04	.05
IL 2	.72	.52	.11	.04
US 1	.30	.89	.39	.05
US 2	.20	.93	.21	.01
KA 1	.45	.66	.25	.09
KA 2	.25	.48	.22	.06
Anteil der Kinder pro Klasse	13.4%	8,9%	42%	35.7%

Innerhalb jeder Klasse ist das Auftreten der Symptome unabhängig, so daß sich die Wahrscheinlichkeit für ein bestimmtes Symptommuster aus dem Produkt der Einzelwahrscheinlichkeiten für die einzelnen Symptome ergibt. So ist z.B. die Auftretens-Wahrscheinlichkeit für das Symptommuster "1 1 0 0 0 0" (nur die beiden Symptome

zum instabilen Leistungsverhalten wurden beobachtet) in den einzelnen Klassen wie folgt zu berechnen:

Klasse 1: .87 x .72 x (1 - .30) x (1 - .20) x (1 - .45) x (1 - .25) = .14469
Klasse 2: .57 x .52 x (1 - .89) x (1 - .93) x (1 - .66) x (1 - .48) = .00040
Klasse 3: .04 x .11 x (1 - .39) x (1 - .21) x (1 - .25) x (1 - .22) = .00124
Klasse 4: .05 x .04 x (1 - .05) x (1 - .01) x (1 - .09) x (1 - .06) = .00161

Man sieht, daß dieses Antwortmuster in Klasse 1 wesentlich häufiger ist als in den anderen Klassen. Das Vorliegen eines solchen Antwortmusters ist also ein diagnostisches Indiz, daß die entsprechende Person aus Klasse 1 stammt.

Berücksichtigt man auch die relative Größe der einzelnen Klassen, so kann man berechnen, mit welcher Häufigkeit das Antwortmuster "1 1 0 0 0 0" bei Modellgeltung in den Daten vorkommen müßte, und diesen Wert mit der empirisch gefundenen Häufigkeit vergleichen. In unserem Beispiel müßte der relative Anteil sich wie folgt ergeben:

p(110000) = .14469 x .134 + .00040 x .089 + .00124 x .42 + .00161 x .357 = .0205

In einer Stichprobe von n = 1172 Kindern ist demnach bei Modellgeltung zu erwarten, daß dieses Symptommuster bei 1172 x .0205 = 24 Kindern beobachtet wird.

Insgesamt liegen für die 64 Antwortpatterns 64 empirische Häufigkeiten vor, denen 64 theoretische gegenüberstehen. Da die Zahl der aus den Daten geschätzten Parameter deutlich kleiner ist als die Zahl der zu erklärenden Häufigkeiten, kann man fragen, ob die empirischen Häufigkeiten von den aus den geschätzten Parametern berechneten, theoretisch erwarteten Häufigkeiten signifikant abweichen. Wenn das der Fall wäre, wäre das Modell zu verwerfen. Bei der vorliegenden Studie waren die Abweichungen nicht signifikant, die Vier-Klassen-Lösung konnte also akzeptiert werden.

Versucht man die vier Klassen inhaltlich zu interpretieren, so hat man die Symptomwahrscheinlichkeiten in den einzelnen Klassen zu vergleichen: Klasse 4 zeichnet sich in allen drei Bereichen durch weitgehende Symptomfreiheit aus. Auch Klasse 2 zeigt noch relativ geringe Symptombelastung und könnte als eine Klasse von Grenzfällen mit leichterer Symptomatik gelten. Die beiden kleineren Klassen 1 und 2 sind beides Klassen mit hoher Symptombelastung, wobei sie sich untereinander durch die Art der Symptome unterscheiden: In Klasse 1 zeigt sich eine besonders hohe Auftretenswahrscheinlichkeit für instabiles Leistungsverhalten, in Klasse 2 eine sehr hohe Auftretenswahrscheinlichkeit für unangemessenes Sozialverhalten. Beide Gruppen können als Gruppen von Problemkindern betrachtet werden.

Die Grundgedanken des Latent-Class-Modells wurden von Lazarsfeld bereits 1950 und nochmals ausführlicher von Lazarsfeld & Henry (1968) dargestellt. Das Modell wurde inzwischen in verschiedener Hinsicht erweitert: Die Items können mehr als zwei Antwortkategorien haben, für die Parameter können verschiedene Restriktionen gesetzt werden (etwa derart, daß bestimmte Klassen gleich groß sein sollen, daß bestimmte Symptomwahrscheinlichkeiten in einer Klasse größer sein sollen als in einer anderen, usw.). Einen Überblick über verschiedene Arten von Latent-Class-Modellen mit unterschiedlichen Arten von Restriktionen findet man bei Formann (1984) und bei Langeheine & Rost (1988). Rost (1988) geht ausführlich auf die formalen Beziehungen zwischen verschiedenen Latent-Trait-Modellen und unterschiedlich restringierten Latent-Class-Modellen ein.

Zusammenfassung

Latent-Trait-Modelle gehen von der Annahme eines latenten Kontinuums (Eigenschaft, Fähigkeit) aus, auf der jede Person einen bestimmten Wert (Personparameter) hat. Die Itemcharakteristik gibt an, wie die Lösungswahrscheinlichkeit (allgemeiner: die Wahrscheinlichkeit für eine bestimmte Antwortkategorie) für ein Item von der Position des Probanden auf dem latenten Kontinuum abhängt. Die Antworten auf die einzelnen Items werden als lokal stochastisch unabhängig vorausgesetzt.

Das Rasch-Modell setzt logistische Itemcharakteristiken voraus. Die Items unterscheiden sich nur in einem Parameter, dem Schwierigkeitsparameter. Besondere Vorzüge sind die Existenz erschöpfender Statistiken (die Trefferzahl ist eine erschöpfende Statistik für den Personparameter) und die spezifische Objektivität. Letztere bildet auch die Grundlage für die statistischen Tests zur Überprüfung der Modellgeltung.

Das Rasch-Modell hat verschiedene Weiterentwicklungen erfahren: (a) Im linear logistischen Modell wird der Schwierigkeitsparameter in additive Komponenten zerlegt. Damit können Hypothesen darüber, wie die Itemschwierigkeiten zustande kommen, überprüft werden. (b) Das mehrkategoriale Rasch-Modell läßt pro Item mehr als zwei Antwortkategorien zu, die geordnet oder bloß qualitativ verschieden sein können. (c) Das Birnbaum-Modell läßt außer dem Schwierigkeitsparameter auch einen Trennschärfeparameter zu; in einer weiteren Variante einen zusätzlichen Parameter für die Wahrscheinlichkeit, bei bloßem Raten das Item zu lösen.

Dem Latent-Trait-Ansatz verwandt ist das LLRA-Modell, das zwar lokale Unabhängigkeit, aber keinen allen Items gemeinsamen latenten Trait annimmt. In formal enger Beziehung zum Latent-Trait-Ansatz steht auch das Latent-Class-Modell, das statt einer quantitativen latenten Dimension eine Einteilung der Probanden in qualitative latente Klassen zum Ausgangspunkt nimmt.

Einführende Literatur:

Kubinger, K.D. (1988). Testtheorie: Probabilistische Modelle. In R.S. Jäger (Hrsg.), ***Psychologische Diagnostik.*** Ein Lehrbuch. (S. 264 - 276). München: Psychologie Verlags Union.

Weiterführende Literatur:

Fischer, G.H. (1974). ***Einführung in die Theorie psychologischer Tests.*** Bern: Huber.

Fischer, G.H. (1983). Neuere Testtheorie. In H. Feger & J. Bredenkamp (Hrsg.), ***Messen und Testen*** (S. 604 - 692). Göttingen: Hogrefe.

Kubinger, K.D. (1988). Aktueller Stand und kritische Würdigung der probabilistischen Testtheorie. In: Kubinger, K.D. (Hrsg.), ***Moderne Testtheorie - Ein Abriß samt neuesten Beiträgen*** (S. 19 - 83). Weinheim: Beltz.

Langeheine, R. & Rost, J. (Eds.) (1988). ***Latent trait and latent class models. New*** York: Plenum Press.

Neuere Anwendungen, bei denen unterschiedliche Modellvarianten und Schätzverfahren an denselben Daten vergleichend erprobt wurden, findet man u.a. bei:

Glas, ***C.*** A. W. (1989). ***Contributions to estimating and testing Rasch-models.*** Dissertation der Universität Twente. Den Haag: CIP-Gegevens Koninklijke Bibliothek. ISBN 90-9003078-6.

Haertel, E.H. (1990). Continuous and discrete latent structure models for item response data. ***Psychometrika, 55, 477 - 494.***

8. Adaptives Testen

1. Was versteht man unter adaptivem Testen?
2. Welchen Beitrag leisten Latent-Trait-Modelle um zu entscheiden, welches Item dem Probanden als nächstes vorgelegt wird?
3. Wie können mit Hilfe von Latent-Trait-Modellen Punktwerte trotz unterschiedlicher Itemauswahl verglichen werden?

Vorstrukturierende Lesehilfe

Zunächst wird auf Vorläufer antwortabhängigen adaptiven Testens hingewiesen. Danach wird gezeigt, daß Latent-Trait-Modelle sowohl auf die Frage nach der für einen Probanden optimalen Bemauswahl als auch auf die Frage der Vergleichbarkeit von Testwerten trotz von Proband zu Proband unterschiedlicher Itemauswahl eine präzise Antwort zu geben vermögen. Schließlich wird auf einige Studien hingewiesen, die über praktische Erfahrungen berichten.

Adaptives oder antwortabhängiges Testen zeichnet sich gegenüber konventioneller Testvorgabe dadurch aus, daß die Auswahl der Testaufgaben, die der Proband zu bearbeiten hat, nicht schon zu Beginn festliegt, sondern erst während der Testdurchführung in Abhängigkeit von den bisher gegebenen Antworten erfolgt. Das entspricht dem Grundkonzept nach dem, was in nicht formalisierter Weise wohl jeder Prüfer tut, der eine mündliche Prüfung abhält: Er wertet laufend die Antworten aus, macht sich ein Bild vom Kenntnisstand des Probanden und modifiziert den Schwierigkeitsgrad seiner Fragen: Wenn der Proband nicht richtig geantwortet hat, wählt er leichtere, wenn er richtig antwortet schwerere Fragen.

Adaptive Strategien wurden auch schon seit der Anfangszeit der Intelligenzmessung, z.B. in den Binet-Tests (siehe z.B. Binetarium nach Norden, 1930) verwendet. Die Aufgaben sind dort der Schwierigkeit nach geordnet und bestimmten Lebensaltem zugeordnet. Das Kind bekommt zunächst Aufgaben gestellt, die der Schwierigkeit nach 1 Jahr unter seinem Lebensalter anzusiedeln sind, und je nach Erfolg oder Mißerfolg bei diesen Aufgaben wird mit Aufgaben höherer oder niedrigerer Altersstufen fortgefahren. Auch die Beendigung erfolgt antwortabhängig: Wenn die Aufgaben einer Altersstufe alle nicht mehr bewältigt wurden, werden keine weiteren Aufgaben mehr gestellt.

Bei den in der Folgezeit entwickelten Tests treten adaptive Verfahrensweisen allerdings nur noch vereinzelt auf: Bei Tests mit Einzeldurchführung werden die Abbruchkriterien gewöhnlich in Abhängigkeit vom Erfolg bzw. Mißerfolg des Probanden festgelegt. So ist z.B. beim HAWIE (Hamburg-Wechsler-Intelligenztest nach Hardesty & Lauber, 1956) zu den einzelnen Subtests jeweils eine Abbruchregel angegeben: Wenn eine bestimmte Anzahl von Items hintereinander nicht gelöst wurde,

so werden die weiteren, schwierigeren Items dieses Subtests nicht mehr vorgelegt. Die Anwendung des Progressiven Matrizen Tests nach Raven, eines nicht verbalen Intelligenztests, sieht vor, bei vermutlich leistungsschwachen Probanden zunächst mit einer leichteren Form, den "Coloured Progressive Matrices" zu beginnen und je nach Erfolg oder Mißerfolg mit der schwierigeren Standard-Version fortzufahren (Raven, 1963). Solche adaptiv verzweigende Elemente in der Testdurchführung sind aber eher die Ausnahme. Bei der ganz überwiegenden Mehrzahl der Tests ist die Durchführung für alle Probanden gleich: Die Aufgaben. werden der Schwierigkeit nach steigend angeordnet und allen Probanden in der gleichen Weise vorgelegt. Der Hauptgrund dafür ist wohl darin zu sehen, daß im Interesse der Testökonomie die meisten Tests Papier-Bleistift-Tests sind, die in Gruppen durchgeführt werden. Eine Gruppendurchführung mit einheitlicher Instruktion und einheitlicher Bearbeitungszeit für alle Probanden läßt eine individualisierte adaptive Aufgabendarbietung praktisch nicht zu.

Ein weiteres Problem ist eher theoretischer Art: Wenn bei adaptiver Testvorgabe jeder Proband andere Aufgaben bearbeitet hat, so sind die Leistungen untereinander schwer zu vergleichen. Die Leistung eines Probanden, der am Anfang Treffer erzielte und daraufhin schwierigere Fragen bekam, die er nicht mehr beantworten konnte, ist offensichtlich höher zu bewerten als die eines Probanden, der am Anfang einige Fehler machte und seine Treffer bei den daraufhin gebotenen leichten Items erzielte. Aber um wieviel höher? Wenn bei adaptiver Itemauswahl eine Vielzahl unterschiedlicher Item-Abfolgen möglich ist und aufgrund adaptiver Abbruchregeln unterschiedliche Itemzahlen geboten wurden, so ist die Frage nach einem gerechten Punktesystem schwer zu beantworten und auf der Basis von bloßen Ad-hoc-Regeln wohl kaum befriedigend zu lösen.

Eine theoretische Grundlage für adaptive Teststrategien, die nicht nur eine rational begründete Itemauswahl ermöglicht, sondern auch eine theoretische Basis für den Vergleich von Testleistungen trotz unterschiedlicher Itemauswahl liefert, wurde erst mit Hilfe der Latent-Trait-Modelle geschaffen. Zunächst muß gezeigt sein, daß alle Items eines Itempools einem bestimmten Latent-Trait-Modell, z.B. dem einfachen Rasch-Modell, genügen, und die Itemparameter müssen bekannt sein. Wenn das der Fall ist, kann man jede beliebige Teilmenge von Items benützen, um für einen Probanden den Personparameter zu schätzen. Damit ist das Problem der Vergleichbarkeit der Testwerte trotz unterschiedlicher Itemauswahl gelöst: Verglichen werden nicht die Trefferzahlen, sondern die -unter Berücksichtigung der Itemparameter (im Falle des Rasch-Modells der Schwierigkeitsparameter) - geschätzten Personparameter.

Auch die Frage, welches Item als nächstes vorgelegt werden soll, läßt sich präzis beantworten: Wenn es das Ziel ist, mit möglichst wenig Items eine möglichst genaue Schätzung des Personparameters zu erhalten, so ist es die optimale Strategie, während der Testdurchführung laufend den Personparameter zu schätzen und als nächstes immer dasjenige Item auszuwählen, das an der Stelle des geschätzten Personparameters bestmöglich diskriminiert. Im Falle des einfachen Rasch-Modells ist das dasjenige Item, das bei diesem Personparameter die Lösungswahrscheinlichkeit 0.5 hat. Je nach Erfolg oder Mißerfolg bei diesem Item wird die Schätzung des Personparameters nach oben oder unten korrigiert und als nächstes ein um den entsprechenden Betrag schwereres oder leichteres Item geboten.

Im theoretischen Idealfall stehen Items beliebiger Schwierigkeitsabstufung zur Verfügung. Stellt man sich weiter vor, man hätte unter Verwendung des linear-logistischen Modells die Itemschwierigkeiten vollständig durch die zur Lösung erforder-

lichen Operationen erklärt (vgl. Kapitel 7.3), so könnte das nächste Item mit dem gewünschten Schwierigkeitsgrad auch vom Computer erzeugt werden. Realistischer ist es, von einer begrenzten Itemmenge auszugehen, aus der dann immer das Item ausgewählt werden kann, das unter den vorhandenen an der entsprechenden Stelle die relativ beste Trennschärfe hat und somit den größten Informationsgewinn über den Personparameter liefert. Zur laufenden Schätzung des Personparameters wurden verschiedene Verfahren vorgeschlagen, die sich danach unterscheiden, ob man Annahmen über die Verteilung der Personparameter in der Population als Vorwissen mit eingehen lassen will (Bayes-Schätzungen) oder nicht (Maximum Likelihood-Schätzungen). (Eine vergleichende Simulationsstudie findet man bei Wild, 1988a).

Über praktische Erfahrungen mit computerunterstütztem, adaptivem Testen liegen erst einzelne Studien vor. McBridge & Martin (1983) wiesen darauf hin, daß trotz der unbestrittenen theoretischen Überlegenheit adaptiver Testverfahren gegenüber der konventionellen Testvorgabe, die bis dahin in der Literatur berichteten praktischen Anwendungen diese Überlegenheit nicht immer bestätigt hätten. Sie schlossen eine eigene Untersuchung an, die beiden Verfahrensweisen möglichst gleich gute Chancen geben sollte. Die Probanden wurden nach dem Zufall auf die beiden Bedingungen adaptive vs. konventionelle Testvorgabe aufgeteilt. Die Testitems ("verbal ability") stammten aus demselben Itempool von 150 Items. Jede Person hatte 2 Testformen (entweder 2 mal adaptive oder 2 mal konventionelle Vorgabe) zu je 30 Items zu bearbeiten. Die Darbietung erfolgte in jedem Fall per Computer. Bei adaptiver Darbietung wurden die Items gemäß dem aktuellen Stand der Schätzung des Personparameters ausgewählt, bei konventioneller Darbietung wurden die Items so ausgewählt, daß sie den gesamten Schwierigkeitsbereich gleichmäßig abdeckten. Außerdem wurde jeder Person als "Kriteriumsmaß" ein umfangreicher Wortschatztest ("word knowledge") vorgelegt.

Im Ergebnis zeigte sich eine bessere Paralleltest-Reliabilität für die adaptive Vorgabe. Der Unterschied war bei einem sehr kurzen Test am deutlichsten (bei 5 Items .78 für adaptive, .58 für konventionelle Darbietung) und glich sich mit zunehmender Testlänge aus (bei 30 Items .92 für adaptive, .89 für konventionelle Darbietung). Bei der Validität (Übereinstimmung mit dem Kriteriumstest) zeigte sich eine nur geringfügige bessere Korrelation der adaptiven Form. In einer Wiederholung der Studie fielen die Ergebnisse noch deutlicher zugunsten der adaptiven Form aus. Ähnliche Ergebnisse, nämlich eine Verbesserung der Meßgenauigkeit bei adaptiver gegenüber konventioneller Testvorgabe, insbesondere bei kleinen Itemzahlen, aber keine oder keine wesentliche Verbesserung der Validität gemessen an Außenkriterien, traten auch in verschiedenen anderen Studien auf (eine Überblicksdarstellung findet man bei Bloxom, 1989).

In der Studie von McBridge & Martin (1983) wurde unter gleicher Testlänge gleiche Itemzahl verstanden. Wild (1988b), die mit einer adaptiven Variante des Matrizen-Tests arbeitete, berichtet allerdings über deutliche erhöhte Itembearbeitungszeiten bei adaptiver Vorgabe. Damit wird der Effizienzgewinn wieder fraglich. Nährer (1988) schlägt vor, die Bearbeitungszeiten in die Auswahlstrategie mit einzubeziehen und die Items so auszuwählen, daß die bestmögliche Genauigkeit bei minimaler Testzeit (statt bisher: Itemzahl) erreicht wird.

Ein weiteres Problem bei adaptiver Testvorgabe besteht darin, daß sich die Item-Parameter durch Lernen während der Testdurchführung verändern können. Inwieweit das der Fall ist, wird natürlich vom Inhalt des Tests und dem Testmaterial abhängen.

Bei einem Wortschatztest wird Lernen während der Testvorgabe vermutlich kaum eine Rolle spielen, im Unterschied etwa zu Aufgaben, die wiederholte Anwendung derselben Operationen (z.B. Anwendung der Hebelgesetze) erfordern. Wenn Lernen während der Testvorgabe eine nicht vernachlässigenswerte Rolle spielt, so bedeutet das, daß der Itemparameter für ein Item nicht feststeht, sondern von der Position abhängt, an der das Item geboten wird. Gittler und Wild (1988) zeigen in einer Simulationsstudie, daß nicht berücksichtigte Lerneffekte zu einem erheblichen Bias bei der Schätzung der Personparameter führen können. Als publizierte Tests für den routinemäßigen Einsatz in der diagnostischen Praxis stehen Testverfahren mit computerunterstützter maßgeschneiderter ("tailored") Testvorgabe noch nicht zur Verfügung. Das Verfahren, das den Ansatz des adaptiven Testens bisher am weitesten realisiert hat, ohne allerdings Computereinsatz zu benötigen, ist das Adaptive Intelligenz-Diagnostikum (AID) von Kubinger & Wurst (1985). Es werden während der Testdurchführung Zwischenauswertungen durchgeführt, die über die weitere Aufgabendarbietung entscheiden (Näheres siehe Beispiel 8.1). Damit ist ein handhabbarer Weg gefunden, adaptive Testvorgabe auch ohne Computer zu realisieren. Allerdings ist Einzeldurchführung durch einen geübten Versuchsleiter erforderlich. Inwieweit sich die Validitätserwartungen und die von den Autoren erwartete Verbesserung der Motivationslage bei den Probanden bestätigen lassen, bleibt noch zu untersuchen.
Über die Entwicklung eines Lerntests mit computerunterstützter adaptiver Testvorgabe berichten Guthke et al. (1991).

Beispiel 8.1: Adaptives Testen ohne Computereinsatz: Adaptives Intelligenz-Diagnostikum AID von Kubinger & Wurst (1985), Untertest 1 "Alltagswissen"

Das AID besteht aus 11 Untertests. Alle Untertests sind nach dem Rasch-Modell konstruiert und nach verschiedenen Kriterien auf Rasch-Homogenität geprüft. Bei 9 der 11 Untertests ist eine adaptiv verzweigende Durchführung vorgesehen, u.a. bei Untertest 1 "Alltagswissen". Durchführung und Auswertung dieses Untertests laufen wie folgt ab:
Dem Probanden werden zunächst 5 Aufgaben vorgelegt. Je nach Zahl der Richtigen, die vom Versuchsleiter während der Testdurchführung festgestellt wird, ist mit einer von 3 weiteren Aufgabengruppen (einer leichteren, einer gleich schweren oder einer schwereren) fortzufahren. Diese zweite Aufgabengruppe besteht wieder aus 5 Aufgaben. Der Versuchsleiter hat die Zahl der Richtigen in diesem zweiten Aufgabenblock festzustellen und je nach Abschneiden des Probanden im zweiten Block mit einer von drei weiteren Aufgabengruppen fortzufahren. Diese dritte Aufgabengruppe besteht wieder aus 5 Items, so daß der Proband insgesamt 15 Items zu bearbeiten hat.
Im Laufe der Testdurchführung sind also zwei Zwischenauswertungen mit anschließender Verzweigung vorgesehen. Bei drei Alternativen pro Verzweigung ergeben sich somit neun Möglichkeiten für die Zusammenstellung des Tests. Rohwert ist die Zahl der gelösten Aufgaben. Da nun derselbe Rohwert je nach Schwierigkeit der Items, die zu bearbeiten waren, Unterschiedliches bedeuten kann, gibt es für jede der neun Möglichkeiten eine eigene Umrechnungstabelle, die dem Rohwert einen geschätzten Personparameter zuordnet. Diese geschätzten Personparameter sind nun vergleichbar, egal auf welche Weise sie erzielt wurden. Allerdings haben geschätzte Personparameter keine unmittelbar anschauliche Bedeutung und sind diagnostisch schwer interpretierbar. Deshalb werden in weiteren Tabellen diesen geschätzten Personenparametern altersstandardisierte T-Werte zugeordnet.

Zusammenfassung

Grundzüge antwortabhängigen, adaptiven Testens findet man sowohl in alltäglichen Prüfungssituationen als auch in frühen Testkonzepten. Eine theoretische Basis wurde durch die Latent-Trait-Modelle geschaffen. Wenn die Items einem bestimmten Latent-Trait-Modell, z.B. dem Rasch-Modell, genügen, so kann man jeweils dasjenige Item auswählen, das an der Stelle des aktuell geschätzten Personparameters die beste Trennschärfe hat. Will man die Leistungen verschiedener Probanden vergleichen, so kann man das anhand der geschätzten Personparameter tun.

Die Anwendung erfordert in der Regel Computereinsatz. Zum Vergleich zwischen konventionellem und computerunterstütztem adaptivem Testen liegen einige Erfahrungsberichte vor, die teilweise eine bessere Reliabilität und Validität der adaptiven Testvorgabe ausweisen, aber auch auf Probleme (Veränderungen durch Lerneffekte während des Testens, verlängerte Bearbeitungszeit pro Item) hinweisen.

Einführende Literatur:

Kisser, R. (1988). Adaptive Strategien. In R.S. Jäger (Hrsg.), ***Psychologische Diagnostik*** (S. 123-130). München: Psychologie Verlags Union.

Weiterführende Literatur:

Bloxom, B. (1989). Adpative Testing: A review of recent results. ***Zeitschriftfür Differentielle und Diagnostische Psychologie,*** 10, 1-17.

Kubinger, K.D. (1986). Adaptive Intelligenzdiagnostik. ***Diagnostica, 32, 330-344.***

Kubinger, K.D. (1988). ***Moderne Testtheorie.*** Weinheim: Psychologie Verlags Union.

Weiss, D.J. (Ed.) (1983). ***New horizons in testing. Latent trait test theory and computerized adaptive testing.*** New York: Academic Press.

9. Spezielle Probleme der Veränderungsmessung

9.1 Formale und inhaltliche Ansätze zur Messung Vonveränderungen

1. Wie lassen sich Veränderungen, z. B. ein Lerngewinn, im Rahmen verschiedener testtheoretischer Modelle quantifizieren, und wie kann man in diesen Modellen speziellere Hypothesen, z.B. über die Art des Lernens, formal ausdrücken und überprüfen?
2. Welche Vorschläge wurden gemacht, um den Gesichtspunkt der Änderungssensitivität schon bei der Testkonstruktion zu berücksichtigen?
3. Was sind Lerntests, und welche Erfahrungen liegen damit vor?

Vorstrukturierende Lesehilfe

Zunächst wird die Frage, ob eine Veränderung stattgefunden hat, von der Frage abgegrenzt, wodurch diese Veränderung zustande gekommen ist, ob z.B. ein bestimmtes Lernprogramm Erfolg gehabt hat. Dann wird die Frage behandelt, wie Veränderungen, insbesondere auch Veränderungen durch Lerneffekte, in verschiedenen testtheoretischen Modellen dargestellt werden können: in der klassischen Testtheorie, im Rasch-Modell, im linear-logistischen Modell und im Latent-Class-Modell (9.1.1).

Stärker von inhaltlichen Gesichtpunkten als von formalen Modellansätzen ausgehend wurden Vorschläge zur Entwicklung spezieller änderungssensitiver Tests gemacht. Dabei soll die Änderungssensitivität eines Items empirisch bestimmt und als Itemselektionskriterium verwendet werden. Diese Vorschläge werden in (9.1.2) diskutiert.

Ein letzter Abschnitt (9.1.3) befaßt sich mit Lerntests, als einem speziellen inhaltlichen Bereich der Veränderungsmessung. Hier wird nicht nur über den methodischen Ansatz, sondern auch über die inhaltlichen Ergebnisse berichtet, die schließlich zu einem Wandel der Fragestellung geführt haben.

Probleme der Veränderungsmessung treten in allen Bereichen Psychologischer und Pädagogischer Diagnostik auf: bei der individuell beratenden Diagnostik, in der angewandten Forschung, z.B. bei der Evaluation von Förderprogrammen, aber auch bei vielen Fragestellungen in der psychologischen Grundlagenforschung. Dabei sind zwei Hauptfragestellungen zu unterscheiden: Die erste Frage richtet sich darauf, ob überhaupt eine Veränderung stattgefunden hat. Daran schließt sich als zweites die Frage an, wodurch diese Änderung zustande kam: Bei einem Lernexperiment z.B. die Fra-

ge, ob es sich bei der Veränderung um Lerneffekte handelt oder vielleicht nur um Effekte der Testgewöhnung; bei der Erprobung eines Programms zur Förderung der kognitiven Entwicklung z.B. die Frage, ob die Verbesserung dem Förderprogramm zuzuschreiben ist oder anderen Einflüssen, denen die Versuchspersonen in dem Zeitintervall ebenfalls ausgesetzt waren.

Zur ersten Frage, wie die Veränderung festzustellen ist, liegen aus der Testtheorie sowohl Beiträge formaler Art wie auch spezifische inhaltliche Beiträge vor. Sie werden im folgenden Kapitel (9.1) dargestellt. Die zweite Frage, also die Frage nach Nachweis, Abgrenzung und Quantifizierung von Behandlungseffekten, stellt sich vor allem im Bereich von Evaluationsstudien und schließt Probleme der Versuchsplanung, insbesondere der experimentellen und quasi-experimentellen Kontrolle, mit ein. Diese Fragen werden im Kapitel über Evaluationsforschung (9.2) angesprochen.

9.1.1 Die Darstellung von Veränderungen im Rahmen verschiedener testtheoretischer Ansätze

9.1.1.1 In der klassischen Testtheorie

Im Rahmen der klassischen Testtheorie kann man Lernen als Zuwachs im wahren Wert ausdrucken. Liegen von einem Probanden vor und nach einem Training zwei Testwerte X_1 und X_2 vor, so kann man zunächst einmal fragen, ob der Zuwachs $X_2 - X_1$ groß genug ist, daß mit hinreichender Sicherheit ausgeschlossen werden kann, daß er nur durch zufällige Meßfehler zustande gekommen ist. Diese Frage läßt sich mit Hilfe der Kritischen Differenz (siehe Kapitel 3.2) beantworten. Wurde die Nullhypothese $(\tau_2 - \tau_1 = 0)$ verworfen, so ist immer noch die Frage offen, ob die Veränderung des wahren Werts tatsächlich auf Lernen zurückzuführen ist oder z.B. auf triviale Testwiederholungseffekte. Hier kann allenfalls die Handanweisung weiterhelfen, wenn darin Angaben zum Ausmaß von Wiederholungseffekten enthalten sind. Ansonsten wird man die Interpretation nur auf inhaltliche Plausibilität stützen können, wonach Lernen die naheliegendste Erklärung für die Veränderung nach dem Training ist, ohne aber den Lerneffekt von anderen Effekten genau abgrenzen zu können.

Ähnlich stellt sich die Situation dar, wenn es sich nicht um die Veränderung einer einzelnen Person, sondern einer Gruppe von Personen, z.B. einer Schulklasse handelt. Liegen zwei Messungen (vor und nach dem Training) vor, so kann zwar festgestellt werden, ob eine Veränderung stattgefunden hat, aber nicht schlüssig belegt werden, wodurch diese Änderung bedingt ist. Eine Quantifizierung des Trainingsgewinns und darauf aufbauende Fragestellungen (Womit hängt der Lerngewinn zusammen? Wer hat vom Training mehr profitiert?) ist nur im Vergleich mit geeigneten Kontrollgruppen möglich (siehe 9.2).

Wenn in der klassischen Testtheorie Lernen als Zuwachs im wahren Wert dargestellt wird, so läßt sich das am leichtesten mit der Vorstellung von einem quantitativ definierten Fähigkeitskontinuum verbinden, auf dem der Proband ein Stück nach oben gewandert ist. Das bedeutet jedoch nicht, daß qualitative Veränderungen auf der Basis der klassischen Testtheorie nicht erfaßt werden könnten: Eine Änderung der Lösungsstrategie, der Erwerb neuer Algorithmen usw. führt dazu, daß sich die Lösungswahrscheinlichkeiten und Lösungszeiten für bestimmte Aufgaben ändern, daß bestimmte Fehlerarten häufiger oder seltener werden, usw. Welche qualitativen Ver-

änderungen über welche quantitativen Indikatoren erfaßt werden können, bedarf allerdings einer inhaltlichen Theorie.

9.1.1.2 Im einfachen Rasch-Modell

Im Rasch-Modell läßt sich der Lernzuwachs eines Probanden als Zunahme des Personparameters auffassen. Diese Zunahme kann geschätzt werden, wenn der Proband vor und nach der Lernphase jeweils eine Testform bearbeitet hat. Diese Testformen brauchen nicht parallel im Sinn der klassischen Testtheorie sein, müssen aber einer gemeinsamen Rasch-Skala entstammen und gemeinsam normiert sein. Um das zu gewährleisten, sollte eine Voruntersuchung stattgefunden haben, bei der die Items der beiden Testformen zugleich (also ohne dazwischenliegendes Lernen) einer Personenstichprobe vorgelegt und auf Rasch-Homogenität geprüft wurden. Um die Personenparameter numerisch vergleichbar zu machen, müssen sie für beide Testformen gleich normiert sein, z.B. beide auf den Mittelwert Null in der gemeinsamen Analysenstichprobe. Wenn das der Fall ist, so kann aus dem Rohwert vor und nach der Lernphase jeweils der Personparameter geschätzt werden und die Differenz als Schätzung des Zuwachses verwendet werden.

An diese Schätzung des Zuwachses schließen sich dann die gleichen Fragen, wie sie auch im Rahmen der klassischen Testtheorie zu stellen waren: Ist der Unterschied groß genug, daß mit hinreichender Sicherheit ausgeschlossen werden kann, daß er nur durch die Ungenauigkeiten bei den Parameterschätzungen zustande kam? Ist die Veränderung durch Lernen zustande gekommen? - Da jedoch in den meisten Untersuchungen, die mit dem Rasch-Modell arbeiten, die Hauptfragestellung auf die Testkonstruktion, insbesonders auf die Modellgeltung und auf die Prüfung von Hypothesen bezüglich der Itemparameter gerichtet war und nicht auf praktische Fragestellungen der individuellen Diagnostik oder auch der Programmevaluation, wurden solche Themen im Rahmen der Latent-Trait-Modelle bislang wenig bearbeitet.

9.1.1.3 Im linear-logistischen Modell

Im Rahmen des linear-logistischen Modells kann Lernen auf unterschiedliche Art dargestellt werden: Rost & Spada (1983) entwickelten eine Systematik von acht unterschiedlich komplexen Lernmodellen, die aber nicht alle gut interpretierbar und aus realistischen Datenmengen schätzbar sind. Im folgenden sollen daher nur die wichtigsten Varianten betrachtet werden:

Das restriktivste Modell ("globales Lernen") sieht vor, daß der Lernzuwachs für alle Personen gleich ist. Eine Verschiebung aller Personparameter um einen konstanten Betrag nach oben kann formal auch so ausgedrückt werden, daß alle Items bei der zweiten Testdurchführung um denselben Betrag leichter geworden sind. Werden dieselben Items vor und nach einer Lernphase bearbeitet und dann beide Testdurchführungen einer gemeinsamen Rasch-Analyse unterzogen, so sollten sie sich erstens als Rasch-homogen erweisen, und es sollte sich zweitens für jedes Item der Schwierigkeitsparameter nach der Lernphase aus dem Schwierigkeitsparameter vor der Lernphase plus einer für alle Items gleichen additiven Konstante ergeben. Letzteres kann im linear-logistischen Modell als Restriktion bei der Schätzung der Itemparameter eingeführt werden, und bei der Prüfung der Modellgeltung darf die Hinzunahme dieser Restriktion zu keiner signifikanten Verschlechterung der Modellanpassung führen.

Das dargestellte Modell globalen Lernens ist allerdings so restriktiv, daß es schwer sein dürfte, Daten zu finden, die diesem Modell genügen. Läßt man die Möglichkeit offen, daß der Lernzuwachs für die einzelnen Items unterschiedlich ist, so erhält man ein Modell itemspezifischen Lernens. Wie beim Modell globalen Lernens müssen die Items aus erster und zweiter Testdurchführung eine gemeinsame Rasch-Skala bilden. Da aber nun der Schwierigkeitsverlust, der durch Lernen eingetreten ist, bei jedem Item anders sein kann, sind auf die Itemparameter keine Restriktionen zu setzen. Da auch in diesem Modell der Lerngewinn als ein Schwierigkeitsverlust der Items ausgedruckt wird, der für alle Personen in gleicher Weise gilt, setzt auch dieses Modell voraus, daß der Lernzuwachs (genauer gesagt: die Lernzuwächse für die einzelnen Items) bei allen Personen gleich ist.

Eine inhaltlich interessante Variante des itemspezifischen Lernens stellt das Modell des operationsspezifischen Lernens dar. Hier wird zunächst für jedes Item festgestellt, welche Operationen (Anwendungen von Regeln, z.B. Hebelgesetze) wie oft angewendet werden müssen, um die Aufgabe zu lösen. Die Itemparameter werden zunächst auf die Schwierigkeit der beteiligten Operationen als Basisparameter (siehe Kapitel 7.3) zurückgeführt. Dabei wird angenommen, daß bei der zweiten Testdurchführung die einzelnen Operationen unterschiedlich stark vom Lernfortschritt profitiert haben, also unterschiedlich stark in ihrer Schwierigkeit reduziert worden sind. Bei der Modellanpassung wird der Lerngewinn (Schwierigkeitsverlust) für die einzelnen Operationen geschätzt und überprüft, ob sich der Schwierigkeitsverlust der einzelnen Items aus dem Schwierigkeitsverlust der beteiligten Operationen ergibt. Anwendungen aus dem mathematisch-naturwissenschaftlichen Bereich findet man u.a. bei Spada (1976) und Rost (1977). Scheiblechner (1972) nimmt Lernen schon im Zuge der Itembearbeitung innerhalb einer einzigen Testdurchführung an.

Wie bereits erwähnt, läßt ein Modell, bei dem Lernen dadurch dargestellt wird, daß bei gleichbleibenden Personparametern die Items bei der zweiten Testdurchführung leichter werden, und zwar um einen für alle Personen gleichen Betrag, keine interindividuellen Unterschiede im Lernfortschritt zu. Um ein Modell zu erhalten, das auch individuelle Unterschiede im Lernfortschritt zuläßt, muß jede Person durch zwei Personparameter, vor bzw. nach dem Lernen, gekennzeichnet werden. Es müssen dann die Vortestdaten für sich genommen und die Nachtestdaten für sich genommen jeweils dem Rasch-Modell genügen, sie lassen sich aber nicht in einem einzigen Rasch-Modell (mit nur einem Personparameter für alle Items) zusammenfassen. Durch einen technischen Trick, bei dem die Person vor und nach dem Lernen als zwei verschiedene Personen behandelt wird, lassen sich auch in diesem Modellansatz Hypothesen über item- bzw. operationsspezifisches Lernen testen.

Zur Illustration dieses Ansatzes wird in Beispiel 9.1 die Untersuchung von Rost (1977) dargestellt. Dieses Beispiel zeigt, wie im Rahmen des linear-logistischen Modellansatzes unterschiedliche Hypothesen über den Lernprozeß ausgedrückt und getestet werden können. Dabei zeigt sich aber auch, daß inhaltliche Fragen und Fragen der Versuchsplanung (Sind Vortest und Nachtest itemweise parallel? Sind demnach unterschiedliche Schwierigkeitsänderungen auf unterschiedlich starke Unterrichtseffekte zurückzuführen?) genauso auftreten und genauso ernst zu nehmen sind wie bei Verwendung klassischer Methoden.

Beispiel 9.1: Überprüfung von Hypothesen über den Lernprozeß im Rahmen des linear-logistischen Modells

Rost (1977) wandte das linear-logistische Modell an, um den Effekt eines Lernprogramms zum Thema "Erkennen von funktionalen Abhängigkeiten zwischen zwei Meßwertreihen" zu analysieren. Die Versuchspersonen hatten zunächst in einem Vortest 20 Aufgaben unterschiedlicher Art zu bearbeiten, wobei jeweils zwei Meßwertreihen geboten wurden und die mathematische Funktion (z.B. Y = 2X + 3, Y = 6O/X, usw.) erkannt werden mußte, nach der Y aus X hervorging. Die Testaufgaben unterschieden sich in der Art der Funktion, in der Art der Darbietung (mit/ohne textliche Einkleidung) und darin, ob die Meßwertreihen der Funktion genau entsprachen oder kleine "Meßfehler" enthielten. Es folgte ein fünfstündiges Trainingsprogramm und danach eine zweite Testdurchführung. Der zweite Test enthielt ebenfalls 20 Items, die zu denen des Vortests "sachstrukturell parallel" waren (zu jedem Item des Vortests gab es ein Item des Nachtests, das ihm in den genannten Konstruktionsmerkmalen entsprach). Daß Vortest und Nachtest dasselbe messen, wurde damit zwar nicht empirisch belegt, aber doch inhaltlich gut begründet.

Im ersten Schritt der Auswertung wurde für Vortest und Nachtest getrennt überprüft, ob die Items jeweils eine Rasch-Skala bilden. In beiden Fallen wurde - trotz kleinerer Abweichungen - das Rasch-Modell als verwendbar angesehen. Indem Vortest und Nachtest getrennt analysiert wurden, wurde erstens nicht vorausgesetzt, daß die Personparameter gleich bleiben bzw. nur um eine für alle Personen gleiche Konstante zunehmen. Damit sind individuelle Unterschiede im Lernzuwachs zugelassen. Es ist zweitens nicht vorausgesetzt, daß die Itemparameter bei der zweiten Testdurchführung denen bei der ersten Testdurchführung (bis auf eine für alle Items gleiche Konstante) entsprechen. Damit ist itemspezifisches Lernen zugelassen.

Im nächsten Schritt wurden dann restriktivere Modelle geprüft: Zunächst wurden die Schwierigkeitsparameter der Items (genauer gesagt: der beiden strukturgleichen Paarlinge) in Vortest und Nachtest verglichen. Damit sollte festgestellt werden, ob nicht auch ein Modell, das einen für alle Items gleichen Lernfortschritt annimmt (also kein itemspezifisches Lernen zuläßt), den Daten gerecht wird. Das war nicht der Fall: Es zeigte sich, daß die Schwierigkeitsparameter der Items in ihrer Relation zu einander im Nachtest anders ausfielen als im Vortest. Daraus wurde geschlossen, daß tatsächlich itemspezifisches Lernen stattgefunden hat. - Eine solche Interpretation setzt freilich voraus, daß die beiden Items, die als strukturgleiche Paarlinge vor bzw. nach dem Training vorgelegt wurden, ohne dazwischenliegendes Lernprogramm genau gleich schwierig gewesen waren. Es bleibt kritisch anzumerken, daß das eine sehr hohe Anforderung ist, die empirisch nicht überprüft wurde, sondern aufgrund der "strukturellen Parallelität" als erfüllt angesehen wurde. Aufgrund der Erfahrung, daß auch bei relativ eng umschriebenen Konstruktionsregeln unterschiedlich schwierige Items entstehen können (siehe Kapitel 7.3). bleiben in diesem Punkt Zweifel offen.

Als nächstes wurde die Frage geprüft, ob die Annahme individueller Unterschiede im Lernzuwachs (pro Person zwei Personparameter, je einer für Vortest und Nachtest) notwendig ist, oder ob nicht auch ein Modell mit einem für alle Personen gleichen Lernzuwachs (nur ein Personparameter für Vor- und Nachtest, Lernzuwachs als für alle Personen gültiges Leichterwerden der Items dargestellt) den Daten gerecht wird. Das zweite Modell, das durch eine gemeinsame Rasch-Analyse von Vor- und Nachtest ausgedruckt wird, zeigte eine signifikant schlechtere Anpassung als das erste (getrennte Rasch-Analysen von Vor- und Nachtest), so daß die Hypothese eines für alle Personen gleichen Lernzuwachses verworfen wurde. Angenommen wurde somit ein Modell, bei dem sich (1) Lernen auf die einzelnen Items unterschiedlich stark auswirkt und (2) individuelle Unterschiede im Lernfortschritt vorhanden sind.

9.1.1.4 Im Latent-Class-Modell

Im Rahmen des Latent-Class-Modells kann Lernen als Übergang von einer latenten Klasse in eine andere dargestellt werden. Dieser Modellansatz bietet sich an, wenn Lernen als stufenweiser Übergang zwischen qualitativ verschiedenen Stadien gesehen wird. Rindskopf (1983) und Bergan & Stone (1985) entwickelten einen formalen Rahmen, in dem sich hierarchisches Lernen (eine Regel kann nur erlernt werden, wenn eine bestimmte andere bereits bekannt ist) und nicht hierarchisches Lernen (zwei Regeln können unabhängig von einander entweder bekannt oder nicht bekannt sein) auf der Basis unterschiedlich restringierter Latent-Class-Modelle darstellen lassen. Dabei werden Personen, die dieselben Regeln beherrschen/nicht beherrschen, jeweils als eine Klasse betrachtet. Eine kurze Beschreibung des mathematischen Modellansatzes findet man bei Langeheine & Van de Pol (1990), eine Programmbeschreibung bei Van de Pol et al. (1989).

Eine andere Art der Anwendung des Latent-Class-Ansatzes auf Lerndaten findet man bei Wiedl, Schöttke & Gediga (1986). Ihr Interesse ist auf individuelle Unterschiede im Lernfortschritt gerichtet. Sie boten Schülern nichtverbale Problemlöseaufgaben (Farbiger Matrizentest nach Raven) dar, wobei beim zweiten Mal eine zusätzliche Verbalisierungsinstruktion (Aufforderung zum "lauten Denken") gegeben wurde. Sie verwendeten eine Latent-Class-Analyse, um verschiedene Schülertypen (gleichbleibend Leistungsstarke, gleichbleibend Leistungsschwache, Leistungsgewinner, spezifische Verbesserte usw.) zu definieren. Da sie bei einer relativ geringen Aufgabenzahl von nur 5 Items 8 latente Klassen erhielten, bleibt abzuwarten, ob dieser Ansatz auch bei größeren Datenmengen zu einer ökonomischen Klassifizierung führt.

Die Darstellung der verschiedenen Modellansätze sollte zeigen, daß es mit Hilfe von Latent-Trait- und Latent-Class-Modellen möglich ist, unterschiedliche Hypothesen über die Art des Lernprozesses (global, itemspezifisch, operationsspezifisch; mit und ohne Annahme von individuellen Unterschieden; Zuwachs auf einem quantitativ definiertem Kontinuum oder Wechsel zwischen qualitativen Klassen) mathematisch zu fassen und zu prüfen. Dabei zeigt sich ein fließender Übergang zwischen Testtheorie, die primär auf individuelle diagnostische Anwendung gerichtet ist, und Allgemeiner Psychologie, die eher grundlagenorientiert nach der Art der Lernprozesse fragt. Ein fließender Übergang besteht auch zur sogenannten Mathematischen Psychologie, die rein allgemeinpsychologisch orientiert verschiedene probabilistische Prozeßmodelle für Lernvorgänge entwickelt hat. Diese Ansätze werden hier nicht referiert, da sie für die Diagnostik bislang nicht zu praktischen Anwendungen geführt haben. Als weiterführende Literatur sei auf Spada & Kempf (1977) verwiesen.

9.1.2 Änderungssensitivität als Gesichtspunkt bei der Testkonstruktion

Die Forderung, Tests so zu konstruieren, daß sie möglichst sensitiv auf Veränderungen reagieren, wurde vor allem von der Klinischen, aber auch von der Pädagogischen Psychologie gestellt. Änderungssensitive Tests seien im Rahmen der Evaluationsforschung erforderlich, um den Erfolg von Fördermaßnahmen oder Therapien sichtbar zu machen, aber auch in der Individualdiagnostik, um den Effekt einer Intervention im Einzelfall zu überprüfen. Speziell in der Klinischen Psychologie wurde zu Recht kritisiert, daß es wenig sinnvoll ist, zur Beurteilung eines Therapieerfolgs Fragebogen zu verwenden, die zwar psychometrisch durchanalysiert und wohl etabliert sein

mögen, deren Fragen sich aber auf weit zurückliegende Ereignisse beziehen oder so allgemein formuliert sind, daß sie den Probanden dazu veranlassen, bei der Urteilsbildung über einen längeren Zeitraum (Monate, Jahre) zu mitteln. Aktuelle Veränderungen können in solchen Meßinstrumenten nicht zum Ausdruck kommen. In diesem Sinn kritisieren z.B. Hartig (1975) und Krauth (1983c) die Verwendung des MMPI, wenn es darum geht, die psychischen Folgen medizinischer Eingriffe zu beurteilen.

Um solcher Kritik Rechnung zu tragen und Veränderungen gezielter zu erfassen, wurden zwei verschiedene Wege beschritten, die Krauth (1983c) im Anschluß an Bereiter (1963) als direkte und indirekte Veränderungsmessung bezeichnet. Bei der direkten Veränderungsmessung soll der Proband selbst das Ausmaß der Veränderung beurteilen ("Ich hatte in den letzten vier Wochen seltener/häufiger Kopfschmerzen als zuvor"). Modifikationen derart, daß es sich nicht um Selbstauskünfte, sondern um Auskünfte anderer (Eltern, Lehrer) über den Probanden handelt, sind leicht vorstellbar. Inwieweit freilich solche direkten Fragen nach der Veränderung den Befragten überfordern und damit in besonderem Maß subjektiven Verzerrungen unterliegen, wie z.B. suggestiven Einflüssen aufgrund des Wissens um die therapeutischen Erwartungen, ist noch nicht geklärt. Die Konstruktion von Fragebogen zur direkten Veränderungsmessung erscheint zwar auch in verschiedenen Bereichen der Pädagogischen Psychologie als möglich, doch liegen bislang publizierte Skalen nur aus dem Bereich der Klinischen Psychologie vor (z.B. Veränderungsfragebogen des Erlebens und Verhaltens von Zielke, 1978; 1980; Zielke & Kopf-Mehnert, 1978; Fragen zu erlebten gesundheitlichen Veränderungen von Krampen & v. Delius, 1981). Als indirekte Veränderungsmessung bezeichnet Krauth (1983c) Verfahren, bei denen zu zwei Zeitpunkten jeweils der Ist-Zustand erhoben wird. Während die direkte Veränderungsmessung nur für Bereiche in Betracht kommt, die als Selbst- oder Fremdeinschätzung mit Fragebogen zu erfassen sind, setzt eine zweimalige Erhebung des Ist-Zustandes keine spezielle Testart voraus und kommt auch für den Leistungsbereich in Betracht. Um einen änderungssensitiven Test zu konstruieren, sollen die Items Probanden vor und nach einer entsprechenden Maßnahme vorgelegt werden, um dann diejenigen Items auszuwählen, die die Veränderung besonders deutlich anzeigen. Dazu wurden verschiedene Indizes vorgeschlagen, die von Krauth (1983c) vergleichend diskutiert wurden.

Wenn jedes einzelne Item zu einem quantitativen Wert führt (Rating-Skalen, Lösungszeiten oder Ähnliches), liegt es nahe, die durchschnittliche Differenz zwischen zweiter und erster Messung zu betrachten:

$$\overline{D} = \overline{X}_2 - \overline{X}_1$$

Es werden diejenigen Items ausgewählt, die z.B. als Effekt eines Unterrichts den größten durchschnittlichen Zuwachs anzeigen.

Dieses einfache Maß ist allerdings nur dann sinnvoll zu interpretieren, wenn (1) alle Items dieselbe Skala verwenden und wenn (2) die Richtung der möglichen Veränderung (hier: Lernzuwachs) als bekannt vorausgesetzt werden kann.

Wenn die Items nicht auf derselben Skala liegen, so daß zahlenmäßig gleiche Differenzen je nach Item eine ganz unterschiedliche Bedeutung haben, kann man versuchen, eine bessere Vergleichbarkeit herzustellen, indem man für jedes Item den Zuwachs in Streuungseinheiten ausdrückt. Zieht man dazu die Streuung bei der ersten Messung heran, so erhält man als Maß für die Änderungssensitivität eines Items den Index SI:

$$SI = D/s_1$$

Wenn die Richtung der Veränderung (Zuwachs oder Abnahme) nicht als bekannt vorausgesetzt werden kann, sondern die Maßnahme bei einem Teil der Probanden eine Zunahme, bei anderen eine Abnahme der Werte hervorrufen kann, so sind die Mittelwertsdifferenz und darauf aufbauende Indizes keine geeigneten Maße, um die Änderungssensitivität auszudrucken. Wenn sich positive und negative Veränderungen die Waage halten, ist die Mittelwertsdifferenz Null, auch wenn der Meßwert jedes einzelnen Probanden sich stark geändert hat. In diesem Fall ist es zweckmäßig, die durchschnittliche quadrierte Differenz zu betrachten:

$$\overline{D^2} = \frac{\Sigma (X_2 - X_1)^2}{n}$$

n = Zahl der Personen

Auch die durchschnittliche quadrierte Differenz kann aus Gründen der besseren Vergleichbarkeit zwischen den Items standardisiert werden. Dazu kann man sie durch die Varianz bei der ersten Messung oder auch durch die Varianz der Differenzen teilen. Man erhält dann die Indizes:

$$SI * = \overline{D^2} / s_1^2$$

und

$$SI ** = \overline{D^2} / s_D^2$$

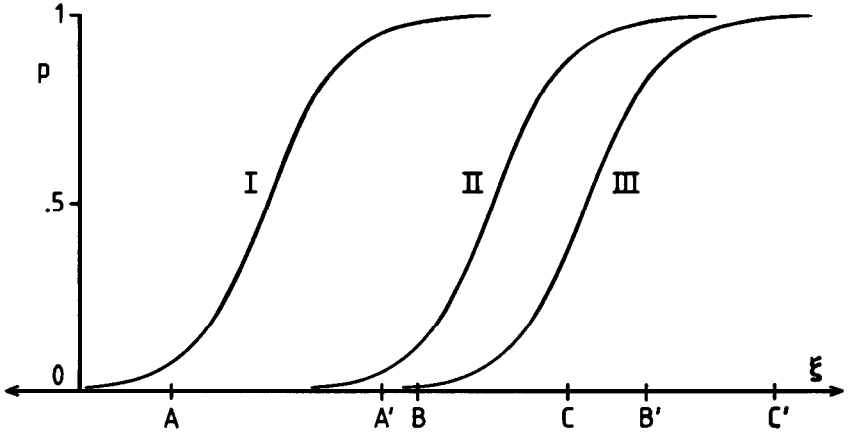
mit s_D^2 = Varianz der Differenzen

Diese Maße der Änderungssensitivität von Items sind zunächst für quantitative Items definiert. Analoge Maße lassen sich auch für Items mit nur zwei Antwortkategorien (richtig/falsch; ja/nein) bilden: Zum einen kann man die Änderung der Item-Schwierigkeit (=Prozentsatz richtiger Lösungen) betrachten. Dieses Maß ist sinnvoll, wenn Änderungen nur in eine Richtung (z.B. Zunahme der Lösungswahrscheinlichkeit) erwartet werden. Wenn Änderungen in beiden Richtungen möglich sind, so kann man den Prozentsatz der Kategorienwechsler berechnen. Weitere Maße, insbesonders auch für Items mit mehr als zwei qualitativ unterschiedenen Antwortkategorien und Fragen der Signifikanzprüfung, sind bei Krauth (1983c) behandelt.

Kritisch ist allerdings anzumerken, daß man wohl nicht erwarten kann, mit Hilfe solcher Indizes zur Änderungssensitivität bestimmte Items ein für alle Male als änderungssensitiv oder nicht änderungssensitiv klassifizieren zu können. Ob und in welchen Items Änderungen auftreten, hängt ja nicht nur vom Inhalt der Items ab, sondern auch von der Art der Maßnahme und der Zusammensetzung der Personenstichprobe.

Die Abhängigkeit von der Zusammensetzung der Personenstichprobe läßt sich an einem einfachen Beispiel demonstrieren: Abbildung 9.1 zeigt drei Items (I,II,II) einer Rasch-Skala und die Positionen dreier Probanden (A, B, C) auf dem Fähigkeitskontinuum. Durch eine Maßnahme (Teilnahme am Unterricht) sei jeder Proband ein Stück auf dem Fähigkeitskontinuum nach rechts gerückt (Positionen A', B', C').

Abbildung 9.1: Unterschiedliche Änderungssensitivität von drei Items (I, II, III) für drei Probanden A, B, C.



Die Änderung des Probanden A von Position A nach A' wird nur von Item I, die Änderung des Probanden B von B nach B' von den Items I und II, die Änderung von Proband C nach C' nur von Item III durch eine große Veränderung in der Lösungswahrscheinlichkeit angezeigt.

Betrachtet man die Lösungswahrscheinlichkeiten für die drei Items vor und nach dem Unterricht, so sieht man, daß die Veränderung des Probanden A vor allem von Item I angezeigt wird (die Lösungswahrscheinlichkeit steigt von nahe Null auf einen Wert nahe Eins), nicht aber von den schwierigeren Items II und III, für die die Lösungswahrscheinlichkeit nach wie vor gering ist. Bei Proband B erscheinen die Items I und II als änderungssensitiv, bei Proband C nur Item III. Je nachdem, ob eine Stichprobe hauptsächlich Probanden vom Typ A, B, oder C enthält (allgemein gesagt: je nachdem, in welchem Skalenbereich sich die Veränderungen abspielen), wird das eine oder andere Item als änderungssensitiver erscheinen. In der Regel will man freilich einen Test nicht auf eine spezielle Stichprobenzusammensetzung hin zuschneiden, sondern ihn so konstruieren, daß Veränderungen in allen Skalenbereichen gut angezeigt werden. Diesem Anliegen entspricht die Empfehlung der klassischen Testtheorie, die Itemschwierigkeiten über den gesamten Bereich zu verteilen, in dem sich Veränderungen abspielen können.

In obigem Beispiel wurde zunächst nur der einfache Fall betrachtet, daß die Items homogen sind und sich die Veränderung formal als Zuwachs auf einem eindimensionalen Fähigkeitskontinuum darstellen läßt. Pädagogische Fördermaßnahmen und Interventionen sind in der Regel komplexer Natur und lassen unterschiedlich starke Wirkungen in verschiedenen kognitiven, emotionalen und motivationalen Bereichen erwarten.

Wenn nun in einem inhaltlich heterogenen Test bestimmte Items oder Itemgruppen keine Veränderung ausweisen, so wäre es kurzschlüssig, diese Items einfach als "nicht änderungssensitiv" auszuscheiden und dem negativen Ergebnis inhaltlich keine Beachtung zu schenken. Wenn z.B. ein Nachhilfeunterricht in Mathematik die Angst vor Klassenarbeiten in Mathematik reduziert, nicht aber die allgemeine Schul-

unlust, so sind beide Ergebnisse pädagogisch relevant. Die Fragen zur allgemeinen Schulunlust als "nicht änderungssensitiv" weil "zu allgemein formuliert" auszuscheiden, hieße einen wichtigen Befund zu ignorieren. Eine andere Maßnahme, die z.B. auf die Verbesserung des Schulklimas insgesamt abzielt, kann gerade umgekehrt bei den Items zur allgemeinen Schulunlust Effekte zeigen, während die Prüfungsangst fachspezifisch gemessen nahezu unbeeinflußt bleibt. Wenn sich dann ein bestimmter Bereich, z.B. eine schwere Verhaltensstörung, bei verschiedenen Maßnahmen immer wieder als kaum beeinflußbar erweist, so ist auch das ein - wenngleich bedauerliches - so doch inhaltlich wichtiges Ergebnis.

Aus den vorgetragenen Argumenten lassen sich folgende Schlußfolgerungen ziehen: Wenn man weder weiß, was die Tests messen noch was die Maßnahme bewirkt, so werden auch Indizes zur Änderungssensitivität nicht viel weiterhelfen. Wer eine Maßnahme evaluieren will, muß Hypothesen darüber haben, worauf sich die Maßnahme auswirkt, und die Tests entsprechend zusammenstellen. Daß sich die Vergangenheit nicht ändern läßt und sich somit Fragen über weit zurückliegende Ereignisse erübrigen, sollte in diesem Zusammenhang trivial sein. Relativ allgemein gehaltene Fragen sind dagegen nicht von vornherein abzulehnen: Ob sich durch eine spezifische Maßnahme, die einen engen Verhaltensbereich betrifft, auch breitere Effekte erzielen lassen, ist in jedem Einzelfall empirisch zu entscheiden, wobei sowohl positive als auch negative Ergebnisse für die Praxis relevant sind.

Abschließend sei nochmals auf einen bereits eingangs betonten Punkt hingewiesen: Weder direkte noch indirekte Veränderungsmessung gibt von sich aus Auskunft darüber, wodurch die Veränderung zustande kam: durch Meßfehler, die zufällig anders ausgefallen sind, durch Testwiederholungs- und Übungseffekte, Reifungsprozesse, Spontanheilung usw., oder eben durch die pädagogische Maßnahme. Alternativerklärungen auszuschalten und eine Interpretation des Effekts als Wirkung der Maßnahme sicherzustellen, ist Sache der experimentellen oder quasi-experimentellen Versuchsplanung. Viele Fragestellungen pädagogischer und psychologischer Evaluationsforschung lassen sich ebenso gut, wenn nicht besser, unter Umgehung der Veränderungsmessung behandeln (siehe Kapitel 9.2).

9.1.3 Der Lerntest-Ansatz

Die Entwicklung von Lerntests stellt einen Ansatz zur Veränderungsmessung dar, der aus einem spezifischen inhaltlichen Anliegen der Pädagogisch-psychologischen Diagnostik entstand. Ursprüngliches Ziel der Lerntestforschung, wie sie im deutschen Sprachraum vor allem durch die Arbeitsgruppe um Guthke (1972) initiiert wurde, war es, die traditionelle Intelligenzdiagnostik, die als bloße Status- oder Zustandsdiagnostik kritisiert wurde, durch die Diagnostik der Lernfähigkeit als einer "Diagnostik intraindividueller Veränderlichkeit" (Guthke 1982), als einer "dynamischen Diagnostik" (Carlson & Wiedl, 1980) zu ersetzen. Dazu sollten in einer standardisierten Lernsituation optimierende Bedingungen geschaffen werden. Die diagnostisch relevante Information sollte dann der Lernfortschritt sein, also der erzielte Zuwachs, nicht der in der Vergangenheit aufgrund hemmender oder fördernder Bedingungen erreichte Zustand. Dadurch sollte speziell bei bisher Benachteiligten vorhandene Lernfähigkeit erkannt werden. Außerdem soll durch die erhöhte "ökologische Validität", die Lerntests in bezug auf Lernanforderungen haben sollen (Guthke, 1982) auch eine bessere prognostische Validität erreicht werden.

Projekte mit ähnlicher Zielsetzung, wie sie in den Sechziger- und Siebzigerjahren in der Arbeitsgruppe um Guthke formuliert wurden, wurden etwa zu gleicher Zeit in verschiedenen anderen Ländern betrieben, so z.B. von Budoff in Cambridge ab 1964, von Feuerstein in Jerusalem ab 1970, von Flammer in der Schweiz 1974. (Näheres dazu findet man bei Kornrann, 1982.)

Im folgenden soll über die Entwicklung von Lerntests nicht nur unter methodischen Gesichtspunkten berichtet werden, sondern es sollen auch die inhaltlichen Ergebnisse mit einbezogen werden, die schließlich zu einem Wandel des Forschungsinteresses geführt haben.

Testmaterial und Vorgehen: Bei der Auswahl des Lernmaterials wird vielfach direkt auf herkömmliche Intelligenztests zurückgegriffen, oder zumindest sehr ähnliches Aufgabenmaterial verwendet. Am beliebtesten sind Aufgaben, bei denen es darum geht, Regeln zu erkennen und anzuwenden oder Beziehungen zu übertragen, also Tests, die nach dem faktorenanalytischen Konzept der Intelligenz hohe Ladungen im Generalfaktor, in Reasoning oder im logisch-induktiven Denken aufweisen. Weitaus am häufigsten wurden Tests vom Muster des Raven-Matrizentests verwendet, weiter Reihenfortsetzungs-Tests (Zahlenreihen, Symbolreihen), Analogieaufgaben (A : B = C : ?), der Mosaik-Test aus dem HAWIE, usw. Im Unterschied zum Vorgehen bei der Intelligenzmessung, wo den Probanden das Material ohne Rückmeldung über die Richtigkeit der Lösung zur selbständigen Bearbeitung überlassen wird, wird bei Lerntests das Lösen der Aufgaben in Interaktion mit dem Versuchsleiter trainiert. Je nach zeitlicher Gestaltung dieses Trainings wird zwischen Kurzzeit- und Langzeit-Lerntests unterschieden: Bei Kurzzeit-Lerntests findet nur eine Testdurchführung statt. Während der Durchführung wird Rückmeldung gegeben und eventuell standardisierte Hilfestellungen geboten. Bei Langzeit-Lerntests findet eine Vormessung statt, daran schließt sich die Unterrichtsphase (Erklärungen, Training) an, danach erfolgt eine zweite Messung. Art und Dauer der Unterrichtsphase kann dabei recht unterschiedlich sein (von 20 Minuten Training zwischen erster und zweiter Testdurchführung bis zu täglichem Training über mehrere Wochen). Eine tabellarische Übersicht über eine Vielzahl von Untersuchungen mit Kurzbeschreibungen des verwendeten Materials und der Art des Trainings gibt Kornrann (1982). Ausführlichere Beschreibungen findet man (außer in den einschlägigen Originalarbeiten) in den zusammenfassenden Darstellungen von Kornrann (1979) oder Guthke (1972; 1980a).

Da bei Kurzzeit-Lerntests nur eine Testdurchführung stattfindet, kann zwischen interindividuellen Unterschieden in der Ausgangslage und im Lernzuwachs nicht unterschieden werden. Als Testwert wird der in dieser einen Testdurchführung erreichte Punktwert verwendet. Bei Langzeit-Lerntests hingegen liegen zwei Messungen vor. Als Maß für den Lerngewinn bietet sich zunächst die Differenz zwischen erster und zweiter Messung an. Solche Differenzen sind aber mit methodischen Problemen (geringe Reliabilität, Skalenprobleme, insbes. Artefakte durch Deckeneffekte) belastet und haben sich auch praktisch nicht bewährt (Guthke, 1972 S. 115; Legler, 1977). Deshalb wird als Testwert durchweg der Wert der zweiten Messung verwendet. Damit wird freilich das ursprüngliche Konzept, Veränderungen zu erfassen, nur unzulänglich realisiert. Insbesonders wird der Sinn der ersten Messung unklar: Ein Vorgehen ohne Vortest, bei dem auf Erklärungen und gemeinsames Training eine Testphase folgt, würde dem gewohnten Schulalltag ebenso gut entsprechen. Flammer & Schmid (1982) weisen zu Recht darauf hin, daß dort, wo es um prognostische Validi-

tät geht, die beiden Messungen mittels multipler Regression optimal zu gewichten wären. Diese optimale Gewichtung dürfte in der Regel weder auf eine alleinige Verwendung der zweiten Messung noch auf eine Differenzbildung hinauslaufen, sondern beiden Testdurchführungen positive Gewichte zuordnen.

Die Bewährung von Lerntests: Zur Bewährung von Lerntests liegt inzwischen eine größere Zahl von Arbeiten, teils einfache Erfahrungsberichte, teils systematisch vergleichende Validitätsstudien vor. Diese wurden bereits in mehreren Übersichtsreferaten gesichtet (Flammer & Schmid, 1982; Guthke, 1972; 1976; 1980a und b, 1982; Guthke & Lehwald, 1984; Kornrann, 1979, 1982; Kornrann & Sporer, 1983).

Am meisten Angaben findet man zum Vergleich zwischen Kriteriumskorrelationen (Schulnoten, Lehrerurteil) von erster und zweiter Messung bei Langzeit-Lerntests. Die erste Messung steht dabei für die konventionelle Diagnostik, die zweite für das Lerntest-Konzept. Die Ergebnisse sind uneinheitlich: In den von Guthke (1972) referierten Arbeiten (überwiegend unveröffentlichte Examensarbeiten aus Leipzig) Enden sich weit häufiger höhere Kriteriumskorrelationen für die zweite Messung als für die erste: Nach einer Auszählung von Flammer (1975) sind bei den Noten als Kriterium in 37 von 41 Stichproben die Korrelationen für die zweite Messung höher (Median der Kriteriumskorrelationen für die zweite Messung 0.60, für die erste 0.49), bei der Intelligenzbeurteilung durch den Lehrer als Kriterium sind sie in 17 von 25 Stichproben höher. Weitere Untersuchungen mit positiven Ergebnissen, insbesondere positive Erfahrungsberichte über Anwendungen im unteren Intelligenzbereich (Hilfsschüler, Debile) sind bei Guthke (1980b) referiert. Dem stehen allerdings negative Befunde anderer Autoren gegenüber: Melchinger (1981) fand in einer Untersuchung mit einem Langzeit-Lerntest (zwischen den beiden Tests lagen 3 Trainings-Sitzungen zu je 2 Stunden) an 175 Schülern/innen der gymnasialen Oberstufe keine höhere Validität des Posttests gegenüber dem Vortest, oder auch des Posttests der trainierten Gruppen gegenüber einer Kontrollgruppe mit bloßer Testwiederholung. Flammer (1974) fand in einem Langzeit-Lerntest (zwischen den beiden Tests lagen zwei Wochen mit täglich einer halben Stunde Training) ebenfalls nur geringe und unsystematische Unterschiede in den Kriteriumskorrelationen (Noten nach dem Übergang zur Oberschule) für erste und zweite Messung. Ähnlich geringe und unsystematische Korrelationsunterschiede fanden Legler (1977) bei Schulanfängern und Wieland (1978, zit. nach Guthke, 1980b) bei Normalschülern (im Unterschied zu fraglich Debilen, für die er positive Ergebnisse berichtet). Insgesamt wird man sich demnach den Schlußfolgerungen von Flammer & Schmid (1982) und Wiedl (1984) anschließen müssen, wonach eine generelle Überlegenheit von Lerntests gegenüber Statustests nicht als belegt gelten kann.

Verschiedene Autoren sind der Frage nach Zusammenhängen zwischen Lerntestergebnissen und möglicherweise leistungshemmenden Persönlichkeitsmerkmalen nachgegangen. Gerade wenn die Lerntestsituation der schulischen Lernsituation stark angeglichen wird, so ist zu vermuten, daß dieselben emotionalen und motivationalen Einflüsse, die den bislang erreichten Schulerfolg determinieren, sich auch in der standardisierten Lernsituation auswirken, was dem Anliegen, kognitive Kapazität zu erfassen, zuwider liefe. Vor allem aus der Leipziger Gruppe (referiert bei Guthke & Lehwald, 1984) liegen eine Reihe von Untersuchungen vor, in denen Fragebogen zur Ängstlichkeit (Testangst, Lernangst), Stress- und Frustrationstoleranz und Neurotizismus mit Lerntestergebnissen korreliert wurden. Die Ergebnisse sind uneinheitlich:

Drei Arbeiten (Stile, 1979; Hentrich & Reich, 1979; Müller 1979; alle zitiert nach Guthke & Lehwald, 1984) berichten über Korrelationen zwischen Ängstlichkeit und erster sowie zweiter Messung bei Langzeit-Lerntests. Entgegen der Erwartung der Autoren waren die Korrelationen zur zweiten Messung nicht niedriger, sondern - sofern signifikante Unterschiede auftraten - höher als zur ersten Messung. Günther & Günther (1981) hingegen fanden bei vier von sechs Lerntests etwas höhere Zusammenhänge zwischen aktueller Befindlichkeit und erster gegenüber zweiter Messung. Weiter sollen Stress- und Frustrationstoleranz etwas höhere Korrelationen zur zweiten als zur ersten Messung zeigen (Guthke & Lehwald, 1984, ohne Quellenangabe).

Bei Kurzzeit-Lerntests fanden Carlson & Wiedl (1976, zit. nach Guthke & Lehwald, 1984) niedrigere Korrelationen des Neurotizismus mit einer Lerntestvariante als mit der Standardversion des Matrizentests. In einer Reihe weiterer Untersuchungen (referiert bei Guthke, 1972; Guthke & Lehwald, 1984) fanden sich keine Korrelationen zwischen Neurotizismus und erster und zweiter Messung in Langzeit-Lerntests.

Insgesamt läßt sich somit wohl nicht belegen, daß Langzeit-Lerntests gerade für ängstliche Personen besonders geeignet waren. Das ist auch verständlich, da ja die zweite Messung im Langzeit-Lerntest ohne Rückmeldung und Hilfen erfolgt, also der schulischen Prüfungssituation gleicht. Für Kurzzeit-Lerntests sieht es möglicherweise anders aus: Wiedl et al. (1982) berichten, daß sowohl eine Kurzzeit-Lerntest-Version (Verbalisierung und Rückmeldung), aber auch eine bloße Verbalisierungs-Instruktion (ohne Rückmeldung) des Raven-Tests, verglichen mit der Standardversion als weniger angstauslösend empfunden wurde. Zumindest bei Einzeldurchführung erscheint es plausibel, daß Verbalisation und Rückmeldung die Testsituation natürlicher und entspannter erscheinen lassen. Bei Kurzzeit-Lerntests, bei denen nur Richtig-Falsch-Rückmeldung gegeben wird, ist jedoch zu bedenken, daß gerade die Leistungsschwächeren viel negative Rückmeldung bekommen, was zu aversiven Reaktionen führen kann (Rollett, 1985).

Verschiedene Untersuchungen befassen sich mit der Frage, ob Unterschiede im kognitiven Stil sich auch bei Lerntests auswirken. Bei Tests vom Typ des Matrizentests erzielen impulsive Kinder schlechtere Ergebnisse als reflexive. Dieses Ergebnis erhält man auch bei Lerntests, sowohl bei Kurzzeit-Lerntests als auch in beiden Messungen bei Langzeit-Lerntests. Dieses Ergebnis wurde in mehreren Untersuchungen bestätigt (Näheres siehe Guthke & Lehwald, 1984). Eine Ausnahme findet man bei Carlson & Wiedl (1980), wo in einer von mehreren Kurzzeit-Varianten die impulsiven Kinder besser abschnitten als die reflexiven.

Wandel des Forschungsinteresses: Wie oben dargestellt, hat der Lerntestansatz die Hoffnungen auf höhere prognostische Validität oder größere Unabhängigkeit von dysfunktionalen emotionalen oder motivationalen Komponenten nicht in befriedigendem Ausmaß erfüllt. Hinzu kommen Forschungsergebnisse, die es fraglich erscheinen lassen, ob sich überhaupt Lernsituationen herstellen lassen, die für alle Schüler gleichermaßen als optimierend gelten können. So z.B. berichten Carlson & Wiedl (1980) zusammenfassend über eine Reihe eigener Untersuchungen, in denen verschiedene Durchführungsarten des Raven-Tests, darunter auch Lerntest-Varianten, mit einander verglichen wurden. Verschiedene Verbalisierungs-Instruktionen (keine Verbalisierung / Vp muß die Lösung begründen / Vp muß auch während des Lösens verbalisieren) wurden mit verschiedenen Rückmeldungsarten (keine / nur richtig oder falsch / richtig oder falsch mit Begründung) kombiniert. Dabei zeigte sich, daß die

Unterschiede zwischen den Durchführungsbedingungen sowohl von der Aufgabenart (Unterteilung des Tests in Aufgabengruppen, Darbietung als Buchform oder als Puzzle) als auch verschiedenen Personmerkmalen (Alter, Leistungsniveau, Impulsivität-Reflexivität) abhingen. Angesichts solcher Ergebnisse erscheint das Ziel, einen Lerntest zu konstruieren, bei dem in einer für alle Probanden optimierenden Lernsituation die wahre Lernfähigkeit zutage tritt, nicht mehr als realistisch.

Die daraus zu ziehende Konsequenz sieht unterschiedlich aus, je nachdem, ob man primär an Grundlagenforschung oder an Anwendung interessiert ist: Interessiert man sich primär für die Prognose des Schulerfolgs, so liegt es nahe, möglichst hohe Übereinstimmung zwischen Lerntest-Situation und schulischer Lernsituation herzustellen. Kornrann (1979) fordert, Lerntests sollten möglichst unterrichtsbezogen sein und in Zusammenarbeit mit Didaktikern aufgrund fächerspezifischer Fehleranalysen entwikkelt werden. Wiedl & Herrig (1978) stellten die Hypothese auf, daß es von der Art des schulischen Unterrichts (konventionell, lehrerorientiert oder "adaptiv", d.h. in Kleingruppen unter Betonung des Verbalisierens und der Selbstkorrektur) abhinge, ob ein Intelligenztest (CFT 1) oder ein Lerntest das Unterrichtsergebnis besser vorhersagt. Die Unterschiede zwischen den Korrelationen gingen in die erwartete Richtung, sind aber (wenngleich inzwischen mehrfach zitiert, z.B. von Flammer & Schmid (1982) als "Nachweis" für die Relevanz "ökologischer Validität") von Signifikanz weit entfernt.

Mehr an Grundlagenforschung interessierte Psychologen (Guthke & Lehwald, 1984) versuchen auf der Grundlage einer Theorie zur allgemeinen Intelligenz näher zu analysieren, welche Teilprozesse durch Training beeinflußt werden. Gegenstand der Prognose sind dann nicht mehr praktisch relevante Validitätskriterien wie Noten oder Lehrerurteil, sondern Leistungen bei gezielt ausgewählten experimentellen Lernanforderungen (Begriffslern-Aufgaben, Mustererkennen), die bestimmte Informationsverarbeitungsprozesse erfordern. Wiedl (1984) weist auf die vielfaltigen Möglichkeiten hin, die sich bei systematischer Variation standardisierter Lernsituationen in verschiedenen Bereichen der Grundlagenforschung (Entwicklungspsychologie einschließlich Alternsforschung, Persönlichkeitspsychologie, Klinische Psychologie usw.) ergeben.

Zusammenfassung

Innerhalb verschiedener psychometrischer Ansätze läßt sich Veränderung auf unterschiedliche Art darstellen: In der klassischen Testtheorie als Zuwachs oder Abnahme im wahren Wert, im Latent-Trait-Ansatz als Zunahme oder Abnahme des Personparameters. Darüber hinaus bieten speziellere Latent-Trait-Modelle die Möglichkeit zwischen globalem, itemspezifischem und operationsspezifischem Lernen zu unterscheiden. Im Latent-Class-Modell kann Lernen als Übergang von einer latenten Klasse in eine andere dargestellt werden.

Weniger von psychometrischen Modellen als von inhaltlichen Fragestellungen ausgehend wurden Vorschläge gemacht, wie man änderungssensitive Tests konstruieren könne: Bei direkter Veränderungsmessung wird der Proband direkt gefragt, ob eine Veränderung aufgetreten ist. Bei indirekter Veränderungsmessung werden die Items zwei Mal vorgelegt und diejenigen Items zu einem änderungssensitiven Test zusammengestellt, die am meisten Veränderung anzeigen. Auf Probleme dieses An-

satzes wurde hingewiesen: Je nach Art der Maßnahme und Zusammensetzung der Probandenstichprobe können jeweils andere Items als besonders änderungssensitiv erscheinen.

Lerntests sind mit dem Ziel entwickelt worden, den Lernfortschritt in einer standardisierten Lernsituation zu erfassen und damit möglicherweise diagnostisch relevantere Information zu erhalten, als das mit einer einmaligen Messung (Messung der Ausgangslage) möglich ist. Die Differenz zwischen Vortest und Nachtest wurde als Maß der Lernfähigkeit schon früh aufgegeben, sowohl aus methodischen Gründen als auch aufgrund mangelnder praktischer Bewährung. Bei Langzeit-Lerntests (Vortest - Lernphase - Nachtest) wurde meist der Nachtest als diagnostisches Maß verwendet; oder aber es findet überhaupt nur eine Testvorgabe statt, bei der durch Rückmeldung und Erklärungen während der Testdurchführung Lernen ermöglicht wird (Kurzzeit-Lerntest). Die Erwartungen, mit Lerntests das Lernpotential unabhängig von der Ausgangssituation bestimmen zu können, und damit insbesonders sozial benachteiligten Kindern besser gerecht werden zu können als mit herkömmlichen Tests, wurden überwiegend nicht erfüllt: Vergleichende Untersuchungen führten zu einer Vielzahl uneinheitlicher Ergebnisse. Als Folge davon trat ein Wandel im Forschungsinteresse auf: Lerntests können zum einen in Richtung auf eine möglichst hohe Übereinstimmung mit der schulischen Lernsituation weiterentwickelt werden, um dann für schulisches Lernen eine möglichst hohe prognostische Validität zu erreichen. Sie können andererseits auch als standardisierte Lernsituationen zu experimentellen Zwecken in der Grundlagenforschung herangezogen werden.

Einführende Literatur:

Petermann, F. (1986). Probleme und neuere Entwicklungen der Veränderungsmessung - ein Überblick. ***Diagnostica, 32, 4-16.***

Weiterführende Literatur:

Zum Problem der Veränderungsmessung allgemein:

Möbus, C. & Nagl, W. (1983). Messung, Analyse und Prognose von Veränderungen. In J. Bredenkamp & H. Feger (Hrsg.). ***Hypothesenprüfung.*** Enzyklopädie der Psychologie, Serie I Forschungsmethoden der Psychologie, Bd.5. (S.239-470). Göttingen: Hogrefe.

zu Kapitel 9.1.1:

Rost, J. & Spada, H. (1983). Die Quantifizierung von Lerneffekten anhand von Testdaten. ***Zeitschrift für Differentielle und Diagnostische Psychologie, 4, 29-49.***

zu Kapitel 9.1.2:

Krauth, J. (1983). Bewertung der Änderungssensitivität von Items. ***Zeitschrift für Differentielle und Diagnostische Psychologie, 4, 7-28.***

zu Kapitel 9.1.3:

Flammer, A. & Schmid, H. (1982). Lerntests: Konzept, Realisierungen, Bewährung. Eine Übersicht. ***Schweizerische Zeitschrift für Psychologie, 41,*** 114-138.

Guthke, J. & Lehwald, G. (1984). On component analysis of intellectual learning ability in learning tests. ***Zeitschrift für Psychologie, 192, 3-17.***

9.2 Methodische Probleme bei der Messung von Behandlungseffekten in der Evaluationsforschung

1. Was sind typische Aufgabenstellungen Pädagogisch-psychologischer Evaluationsforschung?
2. Wie kann bei experimentellem Vorgehen der Effekt einer Maßnahme nachgewiesen werden?
3. Wie lassen sich individuelle Unterschiede im Zuwachs erfassen, und welche methodischen Probleme treten insbesonders im Umgang mit Nachtest-Vortest-Differenzen auf?
4. Welche typischen Probleme treten in quasi-experimentellen Versuchsplänen beim Nachweis von Behandlungseffekten auf?
5. Kann Evaluation auch ohne wissenschaftliche Methodik betrieben werden?

Vorstrukturierende Lesehilfe

Hauptanliegen Pädagogisch-psychologischer Evaluationsforschung ist es, die Wirkung von Maßnahmen, z.B. neuen Förderprogrammen, nachzuweisen und zu analysieren. Wie in 9.2.1 dargestellt, kann dabei die Entwicklung der Fragestellung je nach den vom Auftraggeber gesetzten Vorgaben in unterschiedlichen Stadien abgebrochen oder vertieft und weitergeführt werden. In 9.2.2 werden typische methodische Probleme pädagogischer Evaluationsforschung an drei Beispielen behandelt. Das Thema des ersten Beispiels "Verbalisieren beim Problemlösen" läßt sich experimentell behandeln, so daß der Nachweis des Effekts keine besonderen Probleme aufwirft. Die daran anschließende Analyse des Effekts (Welche Probanden haben vom Verbalisieren mehr, welche weniger profitiert?) ist methodisch schwieriger zu beantworten. An diesem Beispiel werden vor allem Probleme im Umgang mit Nachtest-Vortest-Differenzen behandelt. Dazu zählen Skalenprobleme, Reliabilitätsprobleme und die negative Meßfehler-Kovarianz zwischen Ausgangswerten und Zuwachs.

Anschließend an die methodische Diskussion der Vortest-Nachtest-Differenz als Veränderungsmaß geht es um die Frage der Abgrenzung des Behandlungseffekts (Verbalisieren) von anderen Veränderungen (z.B. durch Gewöhnung und Übung). Dazu ist eine Kontrollgruppe erforderlich. Als Maß des individuellen Behandlungseffekts kann dann die Abweichung von der Regressionsvorhersage aus der Kontrollgruppe verwendet werden. Vor- und Nachteile dieses Maßes werden diskutiert.

Die folgenden beiden Beispiele "Frühförderung der kognitiven Entwicklung" und "Vergleich der Effektivität von Sonderschule und Regelschule bei leistungsschwachen Kindern" dienen der Diskussion von Problemen, wie sie für quasi-experimentelle Forschung charakteristisch sind. Dazu zählen Regressionseffekte, Probleme bei der Zusammenstellung der Kontrollgruppe und selektiver Ausfall von Versuchspersonen.

In einem letzten Punkt (9.2.3) geht es um die These, der Einsatz traditioneller Methodik sei in der Evaluationsforschung überflüssig und durch ein "naturalistisches" Vorgehen zu ersetzen. Aufgrund der in den vorangehenden Abschnitten dargestellten

methodischen Probleme und Fehlerquellen, die oft nicht ohne weiteres erkennbar sind (wie z.B. der Regressionseffekt), dürfte offenkundig sein, wie naiv es ist, diese Probleme mit "naturalistischem" Vorgehen und freier Beschreibung umgehen zu wollen.

9.2.1 Das Anliegen

Neben der individuellen Diagnostik im Rahmen von Beratungssituationen ist die Evaluationsforschung ein weiterer wichtiger Einsatzbereich Pädagogisch-psychologischer Diagnostik. Evaluationsforschung ist primär anwendungsorientierte Forschung, häufig als abgegrenzter Forschungsauftrag von einem Auftraggeber (z.B. einem Ministerium) veranlaßt. Es kann z.B. um die Beurteilung des Erfolgs gezielter Fördermaßnahmen (z.B. Zusatzunterricht bei Lese-Rechtschreibschwäche) gehen, um den Vergleich von Schulsystemen (Gesamtschule versus traditionell dreigliedriges Schulsystem), um die Effizienz von Institutionen (z.B. der Berufsberatung), aber auch um allgemeinere Fragen wie den Vergleich von Unterrichtsmethoden und Lehrstilen.

Ähnlich wie bei der individuellen Diagnostik ist zunächst das Anliegen des Auftraggebers in eine Fragestellung bzw. ein Bündel von Fragestellungen umzusetzen. Dabei kann zunächst eine Beschreibung des Ist-Zustandes im Vordergrund stehen, um auf dieser Grundlage Umfang und Ausmaß des Problems zu beurteilen, z.B.: Wie häufig ist Schulversagen in der Grundschule? Welche Kinder sind betroffen? Wie sieht die weitere schulische und außerschulische Entwicklung dieser Kinder aus?

Aufgrund der Problemanalyse können entweder erste praktische Konsequenzen gezogen werden, oder es können zumindest Hypothesen gebildet werden, mit welchen Maßnahmen (bzw. Änderungen an vorhandenen Maßnahmen) Verbesserungen erzielt werden könnten, auch wenn diese Hypothesen erst noch der empirischen Überprüfung bedürfen. Das kann zunächst in einem Probelauf geschehen, bei dem die Maßnahme weiterentwickelt und evaluiert wird. Auch bei der Evaluation des Probelaufs wird zunächst eine Beschreibung des Ablaufs gefragt sein: Wurde die Zielgruppe erreicht? Konnten die Beteiligten (Kinder, Eltern, Lehrer) zur Mitarbeit gewonnen werden? Wie lief das Programm ab? Welche Probleme traten auf? Wurde die Maßnahme wie geplant zu Ende geführt? - Wenn die Antwort auf diese Fragen zufriedenstellend ausfallt, so schließt sich daran als nächstes die Frage, welche Veränderungen (im Sinne der Zielsetzungen des Programms oder auch positiver wie negativer Nebenerscheinungen) aufgetreten sind und ob bzw. inwieweit diese Veränderungen auf das Programm zurückzuführen sind.

Während bei der Beschreibung der Maßnahme Objektivität, Neutralität und Vollständigkeit der Berichterstattung als methodische Qualitätsanforderungen im Vordergrund stehen, treten bei der Schätzung der Programmeffekte Fragen der versuchstechnischen Kontrolle hinzu, um alternative Erklärungsmöglichkeiten für aufgetretene Veränderungen auszuschließen.

An die Schätzung der Programmeffekte anschließend kann die Fragestellung in verschiedene Richtungen hin weiterentwickelt werden:

(a) Es kann entweder - die Verallgemeinerbarkeit der Ergebnisse voraussetzend - eine Kosten-Nutzen-Analyse bei einer Einführung auf breiterer Basis erstellt werden.

(b) Oder man kann, eingedenk dessen, daß jede Erprobung unter speziellen Rahmenbedingungen stattfindet, vorsichtiger sein und zunächst die Verallgemeinerbar-

keit ausloten, indem man die Maßnahme an verschiedenen anderen Standorten wiederholt. Man wird dann auf eine möglichst genaue Dokumentation des Ablaufs Wert legen, um bei unterschiedlichem Erfolg Hypothesen über die Gründe für Erfolg oder Mißerfolg aufstellen zu können.

(c) Statt ein solches pragmatisch induktives Vorgehen zu wählen, bei dem man - wohl im wesentlichen mit Erfolg rechnend - ausprobiert und nötigenfalls im nachhinein differenzierende Hypothesen aufstellt, kann man auch hier von vornherein vorsichtiger sein und stärker grundlagenorientiert vorgehen. Man wird dann zunächst im hypothesengeleiteten deduktiven Verfahren nach den für den Erfolg entscheidenden Bedingungen suchen. Das betrifft sowohl die Komponenten des Programms, die dann in entsprechenden Kontrollgruppen-Plänen systematisch variiert werden, als auch die Frage, von welchen Eigenschaften der Teilnehmer der Programmerfolg abhängt. Auch diese Fragen bedürfen, wie schon die Schätzung des Programmeffekts, sorgfältiger methodischer Planung.

Welchen Weg Evaluationsforschung geht, hängt nicht zuletzt von den Vorgaben des Auftraggebers ab. Der Auftraggeber kann sich mit einer überwiegend deskriptiv gehaltenen Problemanalyse zufrieden geben, um seine weiteren Entscheidungen nach eigenem Ermessen zu treffen. Oder er kann primär am Ablauf einer von ihm finanzierten Maßnahme interessiert sein, um zu erfahren, was mit seinem Geld geschehen ist. In solchen Fällen wird die mögliche Entwicklung der Fragestellung relativ früh abgebrochen, da der erteilte Auftrag erfüllt ist. Wenn hingegen die Zielsetzung einen weiten Spielraum läßt (z.B. Förderung der Didaktik in den Naturwissenschaften) und der institutionelle Rahmen eine längerfristige Perspektive ermöglicht, ist eine stärker grundlagenorientierte Forschung möglich, deren Ergebnisse dann in einem breiten Bereich anwendungsbezogen nutzbar gemacht werden können. Im folgenden sollen einige typische Probleme Pädagogisch-psychologischer Evaluationsforschung an drei Beispielen erläutert werden. Dabei geht es als zentrale Frage zunächst um den Nachweis eines Effekts, dann aber auch um die weitergehende Frage, wovon der Effekt abhängt. Beim ersten Beispiel handelt es sich um eine Fragestellung aus der Grundlagenforschung, bei der experimentell gearbeitet werden kann. Bei den anderen beiden Fragestellungen stehen Probleme der quasi-experimentellen Kontrolle im Vordergrund.

9.2.2 Beispiele

Beispiel 1: Verbalisieren beim Problemlösen (Probleme im Umgang mit der Vortest-Nachtest-Differenz)

Wir nehmen an, jemand wolle untersuchen, ob Verbalisieren beim Problemlösen ("lautes Denken") die Leistung bei Problemlöseaufgaben verbessert. Dazu kann man einen einfachen experimentellen Versuchsplan verwenden: Die Versuchspersonen werden nach dem Zufall auf zwei Gruppen aufgeteilt, wovon die eine mit, die andere ohne Verbalisieren während des Problemlösens arbeitet. Der Mittelwertsunterschied zwischen den beiden Gruppen kann zur Schätzung des Effekts des Verbalisierens herangezogen und, z.B. mithilfe des t-Tests, auf Signifikanz geprüft werden. Soweit es sich also um den Nachweis des Effekts handelt, werfen Versuchsplan und Auswertung keine besonderen Probleme auf.

Wenn nun als erstes Ergebnis vorliegt, daß sich das Verbalisieren positiv auf das Problemlösen auswirkt, so schließen sich weitere Fragen an. Eine typische Art von Fragestellungen ist darauf gerichtet, zu untersuchen, von welchen Merkmalen des Probanden der Effekt einer pädagogischen Maßnahme abhängt. Im vorliegenden Fall könnte man z. B. fragen, ob Kinder mit hoher versus niedriger Ausgangsleistung, hohem versus niedrigem IQ usw. mehr vom Verbalisieren profitieren. Diese Frage ist allerdings mit dem vorliegenden einfachen Versuchsplan nicht bearbeitbar: Durch den Vergleich von zwei unabhängigen Gruppen läßt sich zwar der durchschnittliche Behandlungseffekt quantifizieren, es läßt sich aber nicht feststellen, ob sich das Verbalisieren individuell unterschiedlich ausgewirkt hat und welche Versuchsperson sich durch das Verbalisieren um wieviel verbessert hat. Infolgedessen hat man auch keine Möglichkeit, den Verbalisierungsgewinn mit anderen Variablen, z.B. dem IQ, zu korrelieren.

Um die Frage beantworten zu können, welche Person wieviel vom Verbalisieren profitiert hat, liegt es nahe, folgenden Versuchsplan zu wählen: Dieselben Versuchspersonen bearbeiten zwei Parallelformen eines Problemlösetests zuerst ohne, dann mit Verbalisierungsinstruktion. Man berechnet für jede Person die Differenz der beiden Testleistungen und korreliert diese Differenzen mit anderen Variablen (Ausgangsleistung, IQ usw.).

Ein solches Vorgehen mag zwar auf den ersten Blick einfach und zielführend erscheinen, enthält aber methodische und inhaltliche Probleme, die im folgenden diskutiert werden sollen. Bei den ersten Punkten (Skalenprobleme, Reliabilität der Differenz, Meßfehlerkorrelation zur ersten Messung) geht es um Fragen, die die Nachtest-Vortest-Differenz als Maß für individuelle Unterschiede in der Veränderung betreffen; danach geht es um die Frage, ob die Veränderung dem Behandlungseffekt (hier: Verbalisieren) gleichgesetzt werden kann.

(a) Skalenprobleme: Die Zahl der gelösten Aufgaben in einem Problemlösetest kann kaum beanspruchen, eine fundierte Intervallskala zu sein. Ob die Skaleneinheiten in verschiedenen Skalenbereichen gleich groß sind, ob z.B. die Differenz zwischen 7 und 9 Richtigen genauso groß ist wie zwischen 17 und 19, läßt sich nicht theoretisch begründet beantworten. Die Frage, ob Kinder mit hohem oder niedrigem IQ mehr vom Verbalisieren profitieren, läuft aber genau auf einen solchen Vergleich hinaus: Die Kinder mit niedrigem IQ haben vermutlich deutlich niedrigere Ausgangswerte als die mit hohem IQ, so daß der Vergleich des Zugewinns einen Vergleich von Differenzen in unterschiedlichen Skalenbereichen erfordert.

In günstigen Fällen, bei sehr drastischen Unterschieden im Zuwachs braucht das nicht problematisch zu werden: Ein günstiger Fall läge z.B. vor, wenn die Gruppe mit den niedrigeren IQ im ersten Durchgang die niedrigere Ausgangsleistung hat, im zweiten Durchgang (mit Verbalisieren) aber dann die Gruppe mit hohen IQ übertrifft. Ein solches Ergebnis bleibt bei monotonen Skalentransformationen (= beliebige Transformationen, bei denen die Reihenfolge der Meßwerte bestehen bleibt) erhalten. Die Aussage “Die Probanden mit niedrigeren IQ haben einen größeren Zuwachs erzielt als die mit hohen IQ” kann hier gemacht werden, auch wenn der Problemlösetest nur Rangskalenniveau hat.

Als nächsten, immer noch günstigen Fall nehmen wir an, die Gruppe mit niedrigen IQ hätte bei Verbalisierungsinstruktion die Gruppe mit hohen IQ zwar nicht übertroffen, sich aber doch von 7 auf 15 Punkte gesteigert, während die mit hohen IQ sich nur von 17 auf 18 verbessert hätte. Falls keine Deckeneffekte vorliegen (von “Dek-

keneffekten" spricht man, wenn ein Test nicht genug schwierige Aufgaben enthält, so daß die besseren Probanden an die maximal erreichbare Punktzahl als "Decke" anstoßen; bei einem anders zusammengesetzten Test hätten sie noch weitere, schwierigere Aufgaben lösen und damit mehr Punkte erreichen können), wird wohl kaum jemand zögern zu sagen, die Gruppe mit niedrigen IQ hätte mehr dazugewonnen als die mit hohen IQ - ungeachtet dessen, daß diese Aussage meßtheoretisch gesehen nicht zwingend ist. Was aber, wenn die untere Gruppe einen Anstieg von 7 auf 12 zeigt (Differenz 5 Punkte), die obere von 17 auf 20 (Differenz 3 Punkte)? Hier kann der Übergang zu einer anderen Skala (z.B. Rangplätzen und darauf aufbauend Rangplatzdifferenzen; Übergang von der Rohwertskala zu den geschätzten Personparametern eines probabilistischen Testmodells) zu einer Umkehr der Interpretation führen, indem einmal für die eine, einmal für die andere Gruppe die Differenz numerisch größer ist. Wenn das der Fall ist, sollte man sich damit begnügen, darzustellen, wie sich die einzelnen Gruppen verbessert haben, aber auf einen numerischen Vergleich der Zuwächse verzichten.

Statt die Kinder nach der Intelligenz in nur zwei Klassen zu teilen (hoher/niedriger IQ) und den Mittelwertsunterschied zwischen den beiden Gruppen zu betrachten, kann man auch einfach die Korrelation zwischen dem IQ und dem Differenzmaß berechnen. Auch das beantwortet die Frage, ob zwischen dem IQ und dem Zugewinn ein Zusammenhang besteht. Was das Skalenniveau anbelangt, gilt dasselbe, was oben im Zusammenhang mit dem Gruppenvergleich angeführt wurde: Wenn Differenzen $(X_2 - X_1)$ mit anderen Variablen (Y) korreliert werden, so sind die Intervalleigenschaften von X kritisch. Eine monotone Skalentransformation von X, die z.B. die Intervalle im unteren Bereich dehnt und im oberen Bereich staucht (oder umgekehrt), kann die Korrelation entscheidend verändern. Um das festzustellen, kann man plausible Skalentransformationen (siehe oben) probeweise durchführen. Wenn das Ergebnis stark variiert, muß man entweder inhaltlich begründen können, warum eine Skala gegenüber den anderen vorzuziehen ist, oder aber auf eine Interpretation der Korrelation verzichten.

(b) Reliabilität: Die Reliabilität einer Differenz $(X_2 - X_1)$ ist meist erheblich niedriger als die Reliabilität von X_1 oder X_2 je für sich genommen. Das soll im folgenden näher begründet werden: Die Varianz einer Differenz besteht aus der wahren Varianz der Differenzen und der Fehlervarianz der Differenzen:

$$\sigma^2 (X_2 - X_1) = \sigma^2 (T_2 - T_1) + \sigma^2 (F_2 - F_1),$$

Die Reliabilität der Differenz ist der Anteil, den die wahre Varianz an der beobachteten Varianz ausmacht:

[9.1] $$\text{Rel } (X_2 - X_1) = \frac{\sigma^2 (T_2 - T_1)}{\sigma^2 (X_2 - X_1)}$$

Wie im folgenden gezeigt wird, hängt dieser Anteil wesentlich von der Korrelation zwischen erster und zweiter Messung ab: Gemäß einem allgemeinen Lehrsatz des Statistik ergibt sich die Varianz einer Differenz von zwei Zufallsvariablen als Summe der Varianzen minus zwei mal der Kovarianz. Wendet man diesen Satz auf die Differenz der Meßfehler an, so erhält man

[9.2] $$\sigma^2 (F_2 - F_1) = \sigma^2 (F_2) + \sigma^2 (F_1) - 2\text{Cov}(F_1F_2) = \sigma^2 (F_2) + \sigma^2 (F_1) - 0,$$

d.h. die Fehlervarianzen addieren sich (die Kovarianz der Meßfehler ist gemäß den Axiomen Null). Dagegen ist bei den wahren Werten die Kovarianz in der Regel nicht gleich Null und von der Summe der Varianzen ist ein entsprechender Betrag abzuziehen:

[9.3] $$\sigma^2 (T_2 - T_1) = \sigma^2 (T_2) + \sigma^2 (T_1) - 2Cov (T_1 T_2)$$

Gewöhnlich korrelieren erste und zweite Messung positiv, so daß die Kovarianz positiv ist. Je höher die Korrelation zwischen erster und zweiter Messung, desto größer die abzuziehende Kovarianz, desto kleiner also die wahre Varianz und damit die Reliabilität der Differenzen. Wenn z.B. bei gleicher Varianz von Vortest und Nachtest jede der beiden Messungen eine Reliabilität von 0.90 hat und Vortest und Nachtest zu .70 korrelieren, ist die Reliabilität der Differenz nur .67, bei einer Vortest-Nachtest-Korrelation von .80 sogar nur .50.

Die niedrige Reliabilität des Differenzmaßes führt dazu, daß Korrelationen dem Betrag nach niedrig ausfallen, selbst dann, wenn zwischen dem meßfehlerfrei gemessenen Zuwachs und dem IQ ein enger Zusammenhang besteht. Dieses Problem ist insofern nicht von grundsätzlicher Bedeutung, als bei bekannter Reliabilität mit Hilfe der Minderungskorrektur (siehe Kapitel 2.2) auch die Korrelation mit den wahren Differenzen berechnet werden kann. Diese meßfehlerbereinigte Korrelation ist allerdings nur von Interesse, wenn es, wie im vorliegenden Beispiel, um theoretische Fragen geht. Wenn es dagegen um praktische diagnostische Anwendungen geht (wie z.B. bei Lerntests in ihrer ursprünglichen Konzeption), so interessiert die Kriteriumskorrelation der beobachteten Differenzen, denn nur diese stehen für die Prognose zu Verfügung.

Reliabilität ist eines der Hauptgütekriterien der klassischen Testtheorie. Das legt ein Mißverständnis nahe: Wenn Differenzen nur eine geringe oder vielleicht gar keine Reliabilität haben, so könnte man meinen, sie seien deshalb nicht geeignet, Veränderungen zu erfassen, inbesondere also auch nicht als Maß für den durchschnittlichen Zuwachs einer Gruppe (etwa im Vergleich zu einer Kontrollgruppe) verwendbar. Daß das ein Mißverständnis wäre, soll im folgenden erläutert werden:

Eine niedrige Reliabilität der Differenzen bedeutet, daß ein großer Teil der Varianz der Differenzen auf Meßfehler zurückgeht. Sie besagt aber nichts über den Mittelwert der Differenzen, also den durchschnittlichen Zuwachs, und die Genauigkeit, mit der er erhoben werden kann. Das wird besonders deutlich, wenn man einen Extremfall betrachtet: Wenn alle Probanden genau den gleichen Zuwachs im wahren Wert haben, so ist die Varianz des wahren Zuwachses Null. Folglich ist die Reliabilität der Differenz, definiert als Anteil der wahren Varianz an der beobachteten Varianz, gleich Null, und die gesamte Varianz der Differenzen ist nur auf Meßfehler zurückzuführen. Trotzdem kann der durchschnittliche Zuwachs als Mittelwertsdifferenz zwischen erster und zweiter Messung berechnet und zur Schätzung des durchschnittlichen (hier zugleich für jeden einzelnen Probanden gültigen) wahren Zuwachses verwendet werden. Diese Schätzung weist sogar eine besonders gute Genauigkeit auf, wie man am Konfidenzintervall sehen kann. Das Konfidenzintervall für $P_1 - \mu_2$ lautet bei $\alpha = .05$:

[9.4] $$(\overline{X}_1 - \overline{X}_2) \pm 1.96\sigma (X_2 - X_1)\ {}^1/\sqrt{n}$$

n = Stichprobenumfang

Es wird umso kleiner, je kleiner die Varianz der Differenzen ist; am kleinsten also, wenn die wahre Varianz der Differenzen Null ist, so daß die beobachtete Varianz der Differenz nur noch aus der Fehlervarianz besteht.

Dieses formale Ergebnis, wonach bei einer Reliabilität der Differenzen von Null der durchschnittliche Zuwachs besonders genau geschätzt wird, mag zunächst paradox erscheinen, läßt sich aber bei näherem Hinsehen auch inhaltlich gut verstehen: Wenn alle Personen genau den gleichen wahren Zuwachs haben, ist es egal, welche Personen gemessen werden. Im Prinzip würde eine Person stellvertretend für alle genügen. Die Ungenauigkeit in der Aussage über den Zuwachs kommt nur durch Meßfehler, nicht durch die Auswahl der Personen zustande. Wenn dagegen der wahre Zuwachs individuell unterschiedlich ist, hängt die Schätzung des durchschnittlichen Zuwachses auch von der zufälligen Auswahl der Personen ab, die Schätzung wird also ungenauer ausfallen. Nur in diesem zweiten Fall, wenn also individuelle Unterschiede im Zuwachs vorliegen, ist die Reliabilität der Differenzen größer als Null, und es macht einen Sinn zu fragen, womit diese individuellen Unterschiede zusammenhängen.

Entsprechendes gilt, wenn der Zuwachs von zwei Gruppen, z.B. einer Experimentalgruppe und einer Kontrollgruppe, verglichen werden soll. Wenn innerhalb jeder Gruppe keine individuellen Unterschiede im wahren Zuwachs bestehen (z.B. in der Experimentalgruppe alle um denselben Betrag zunehmen; in der Kontrollgruppe bei keiner Person ein Zuwachs auftritt), ist innerhalb jeder Gruppe die Reliabilität der Differenzen Null. Trotzdem kann für jede Gruppe der Mittelwert der Differenz als Schätzung des Zuwachses berechnet und die beiden Gruppen verglichen werden. Lediglich die Frage, womit individuelle Unterschiede im Zuwachs zusammenhängen, gibt auch hier keinen Sinn.

(c) Negative Korrelation zur ersten Messung: Wenn untersucht wird, von welchen Merkmalen der Person der Zuwachs abhängt, wird gewöhnlich auch die Frage gestellt, ob der Zuwachs mit den Ausgangswerten korreliert. Es liegt nahe, diese Frage zu beantworten, indem man die Korrelation zwischen Differenz und erster Messung, also $r(X_1, X_2 - X_1)$ berechnet. Dabei tritt allerdings ein Artefakt auf, das durch Meßfehler in X_1 bedingt ist. Dieses Artefakt erkennt man, wenn man die beiden Maße X_1 und $X_2 - X_1$ jeweils in wahren Wert und Meßfehler zerlegt:

$$X_1 = T_1 + F_1$$

$$X_2 - X_1 = T_2 - T_1 + F_2 - F_1$$

Man sieht, daß in beide Maße der Meßfehler von X_1 eingeht, und zwar mit entgegen gesetztem Vorzeichen. Die Kovarianz von X_1 mit $X_2 - X_1$ ist also

$$[9.5] \qquad \mathrm{Cov}[X_1, (X_2 - X_1)] = \mathrm{Cov}[(T_1 + F_1), (T_2 - T_1 + F_2 - F_1)]$$

$$= \mathrm{Cov}[T_1, (T_2 - T_1)] - \mathrm{Var}(F_1).$$

Selbst wenn Ausgangswerte und Zuwachs, meßfehlerfrei gemessen, unabhängig sind, also $\mathrm{Cov}[T_1, (T_2 - T_1)] = 0$ gilt, tritt bei den beobachteten Werten eine negative Kovarianz und damit eine negative Korrelation zwischen Ausgangswerten und Zuwachs auf.

Das Problem ist allerdings insofern nicht von grundsätzlicher Bedeutung, als bei bekannter Reliabilität von X_1 Korrekturformeln zur Verfügung stehen, mit denen für die negative Kovarianz durch Meßfehler in X_1 korrigiert werden kann (Harris, 1963).

Die Frage nach der Korrelation zwischen Ausgangswerten und Zuwachs hat nur Sinn, wenn erste und zweite Messung auf derselben Skala (z.B. einer Rohpunktskala) erfolgen. Die Varianz der zweiten Messung kann dann größer, kleiner oder gleich groß sein wie die der ersten Messung:
Bezeichnet man den Zuwachs mit Z, also:

$$Z = X_2 - X_1,$$

so erhält man

$$X_2 = X_1 + Z,$$

und für die Varianz der zweiten Messung:

$$\sigma^2(X_2) = \sigma^2(X_1) + \sigma^2(Z) + 2\mathrm{Cov}(X_1 Z). \qquad [9.6]$$

Sind z.B. Ausgangswert und Zuwachs unabhängig, so ist die Varianz von X_2 um die Varianz des Zuwachses größer als die Varianz von X_1.

Werden hingegen erste und zweite Messung nicht auf derselben Skala gemessen, sondern jeweils für sich standardisiert (wie z.B. der IQ auf jeder einzelnen Altersstufe), so verliert die Frage nach der Korrelation zwischen Ausgangswert und Zuwachs ihren Sinn. Aus der Formel [9.6] sieht man, daß $\sigma^2(X_1) = \sigma^2(X_2)$ nur gelten kann, wenn die Kovarianz zwischen Ausgangswert und Zuwachs negativ ist (oder wenn die Varianz des Zuwachses Null ist). Deshalb ist z.B. die Korrelation des IQ mit 6 Jahren (Messung X_1) mit der IQ-Änderung von 6 nach 8 Jahren ($Z = X_2 - X_1$) als Folge der altersspezifischen Standardisierung in einer repräsentativen Stichprobe zwangsläufig negativ, wobei sich der Betrag der Korrelation allein aus der Kenntnis der Korrelation der IQ zu den beiden Zeitpunkten ($r(X_1 X_2)$) errechnen läßt (eine Formel dazu findet man bei Stelzl, 1982, S.214; eine inhaltliche Diskussion zur Frage der Korrelation zwischen Ausgangswert und Zuwachs bei der Entwicklung der Intelligenz findet man bei Merz & Stelzl, 1973).

(d) Die Abgrenzung des Behandlungseffekts gegen andere Veränderungen und der Residualgewinn als Alternative zum Differenzmaß: Bisher wurde das Differenzmaß als Maß zur Erfassung von Veränderungen diskutiert. Dabei wurde noch offen gelassen, wodurch die Veränderung herbeigeführt wurde. Gerade bei dem Beispiel "Effekt des Verbalisierens auf das Problemlösen" wäre bei zweimaliger Testung derselben Personen (erst ohne, dann mit Verbalisieren) auch an Effekte der Gewöhnung an die Testsituation und an Übungseffekte zu denken. Es ist auch nicht unplausibel, anzunehmen, daß dieser Gewinn durch Gewöhnung und Übung individuell unterschiedlich ist und mit dem IQ korreliert. In der Nachtest-Vortest-Differenz sind diese Effekte mit den Verbalisierungseffekten vermengt, so daß ohne Hinzunahme weiterer Information keine klare Interpretation möglich ist.

Das Problem, daß die Veränderung auf eine Vielzahl möglicher Ursachen zurückgehen kann, ist für einfache Vorher-Nachher-Versuchsplane typisch und wird auch am folgenden Beispiel (Frühförderung der kognitiven Entwicklung) illustriert. Ein großer Teil der Probleme kann in der Regel gelöst werden, wenn man eine geeignete Kontrollgruppe zur Verfügung hat.

Sind Versuchsgruppe und Kontrollgruppe nach dem Zufall gebildet, so kann man die Versuchsgruppe erst ohne, dann mit Verbalisieren, die Kontrollgruppe beide Male ohne Verbalisieren arbeiten lassen. Die naheliegendste Auswertung besteht darin, den durchschnittlichen Zuwachs (Nachtest-Vortest-Differenz) für beide Gruppen zu vergleichen. Hat die Versuchsgruppe einen größeren Zuwachs erzielt als die Kontrollgruppe, so kann das dem Verbalisieren zugeschrieben werden. Man kann in Anschluß an dieses Ergebnis die Gesamtgruppe unterteilen (z.B. Personen mit niedrigem versus hohem IQ) und für die beiden Teilgruppen die Zuwächse in Versuchsgruppe und Kontrollgruppe vergleichen. Die Kontrollgruppe dient in jedem Fall dazu, Veränderungen, die auf das

Verbalisieren zurückzuführen sind, gegen Veränderungen abzugrenzen, die auf bloße Testwiederholung zurückgehen.

Die Daten der Kontrollgruppe können aber auch anders verwendet werden: Man kann daraus eine Regressionsgleichung berechnen, mit der die Testwerte bei der zweiten Testdurchführung (ohne Verbalisieren) aus den Testwerten der ersten Testdurchführung geschätzt werden. Diese Regressionsgleichung wird dann in der Versuchsgruppe verwendet, um für jede Person aufgrund ihrer ersten Testleistung zu schätzen, welchen Wert sie bei der zweiten Testdurchführung ohne Verbalisieren erzielt hätte. Die Abweichung ihres in der Verbalisierungsbedingung tatsächlich erreichten Testwertes von diesem Schätzwert, der sogenannte ***Residualgewinn,*** wird dann als Maß für den Verbalisierungsgewinn herangezogen. Man kann dann zum einen fragen, ob der durchschnittliche Residualgewinn größer als Null ist; man kann weiter fragen, mit welchen Eigenschaften der Person er zusammenhängt, ob z.B. Personen mit hohem IQ im Durchschnitt einen höheren Residualgewinn aufweisen als Personen mit niedrigem IQ.

Der Hauptvorteil einer solchen Auswertung liegt darin, daß für erste und zweite Messung keine Paralleltests zur Verfügung zu stehen brauchen. Eine Regressionsschätzung verlangt nicht, daß die Skaleneinheiten der beiden Messungen irgendwie vergleichbar sein müßten (aus der Körpergröße in Zentimetern kann das Gewicht in Kilogramm geschätzt werden). Da exakte Parallelität von Testformen schwer zu erreichen ist, ist es immer ein Vorteil, wenn man auf eine solche Voraussetzung verzichten kann. Dafür nimmt man Unsicherheiten in Kauf, die mit der Schätzung der Regressionsgleichung verbunden sind. Bei einem kleinen Stichprobenumfang in der Kontrollgruppe kann diese Unsicherheit beträchtlich sein. Welche Auswertungsart vorzuziehen ist, bleibt im Einzelfall zu entscheiden.

Beispiel 2: Frühförderung der kognitiven Entwicklung (Regressionseffekte, Probleme quasi-experimenteller Kontrolle)

Wie in 9.2.1 herausgestellt wurde, sind die Kernfragen der Evaluationsforschung (1) ob ein Behandlungseffekt nachweisbar ist und (2) wovon dieser Effekt abhängt. Im vorangehenden Beispiel war die erste Frage experimentell entscheidbar. Erst die zweite Frage, bei der es um individuelle Unterschiede im Behandlungseffekt ging, warf methodische Probleme auf (Maß für die individuelle Veränderung, Abgrenzung des Behandlungseffekts gegen andere Veränderungen).

Häufig sind allerdings in der Evaluationsforschung aufgrund praktischer Gegebenheiten experimentelle Bedingungen überhaupt nicht herstellbar, so daß auch die erste Frage (Nachweis und Quantifizierung des Behandlungseffekts) nicht mit einfachen experimentellen Versuchsplänen zu lösen ist. Stattdessen müssen dann quasi-experimentelle Anordnungen und Methoden der Feldforschung herangezogen werden. Auf typische Probleme, die dabei auftreten, haben Campbell & Stanley (1963) in einem viel beachteten Aufsatz hingewiesen. Eine umfassendere Darstellung findet man u.a. bei Cook & Campbell (1979), eine ausführliche methodische Diskussion detailliert dargestellter Forschungsprojekte u.a. bei Cronbach (1983). Im folgenden sollen anhand von zwei Beispielen typische Probleme nicht-experimenteller Forschung diskutiert werden. Bei dem zunächst dargestellten Beispiel, das einer experimentellen Fragestellung noch relativ nahe kommt, wird vor allem auf Unzulänglichkeiten eines einfachen Vorher-Nachher-Versuchsplans und die Notwendigkeit einer Kontrollgruppe hingewiesen. Dabei wird der Regressionseffekt ausführlicher behandelt, da er als Fehlerquelle bei der Interpretation von Vorher-Nachher-Plänen oft nicht leicht zu durchschauen ist.

Wir nehmen an, ein sich über mehrere Monate erstreckendes kognitives Trainingsprogramm für Vorschulkinder aus sozial stark benachteiligten Stadtteilen sollte erprobt werden (Kurzbeschreibungen realer Projekte dieser Art findet man bei Bronfenbrenner, 1974; ausführlichere Darstellungen u.a. bei Zigler & Valentine, 1979). Die Hauptfragestellung des Projekts wäre: Um wieviel haben sich die Kinder durch das Förderungsprogramm verbessert? Daran könnten sich (analog zum Beispiel "Verbalisieren beim Problemlösen") als weitere Fragen anschließen: Welche Kinder haben aus dem Programm am meisten Nutzen gezogen? Welche Komponenten des Programms sind für den Erfolg entscheidend?

Unzulänglichkeiten eines Vorher-Nachher-Versuchsplans: Die Beantwortung zumindest der Hauptfragestellung (durchschnittlicher Effekt des Programms) mag zunächst einfach erscheinen: Man wählt bedürftige Kinder für das Projekt aus, führt zu Beginn des Projekts eine Eingangsmessung (Breitband-Diagnostikum kognitiver Fähigkeiten) und nach Ende des Programms eine zweite Messung durch. Der Erfolg des Programms sollte sich am Unterschied der beiden Messungen (Mittelwertsdifferenz zwischen Vortest und Nachtest) zeigen.

So einfach dieser Versuchsplan auch aussieht, so führt er in der Regel doch nicht zu schlüssigen Ergebnissen. Wie schon am Beispiel "Verbalisieren beim Problemlösen" ausgeführt, enthält die Vortest-Nachtest-Differenz nicht nur die Effekte des Programms, sondern auch andere Komponenten. Da wären einmal triviale Effekte der Testwiederholung (bei den meisten Intelligenztests sind Übungsgewinne auch nach einem längeren Zeitraum noch nachweisbar), der Vertrautheit mit dem Versuchsleiter, der Testsituation usw. Wenn zu den beiden Zeitpunkten verschiedene Tests verwendet werden, können unterschiedliche Verzerrungen in den Eichdaten der beiden Tests dieselbe Probandengruppe einmal etwas günstiger, einmal etwas schlechter abschneiden lassen (Problem der skalenmäßigen Vergleichbarkeit von erster und zweiter Messung).

Ein methodisches Artefakt, das weniger leicht zu erkennen ist, ist der Regressionseffekt aufgrund einer Selektion nach der ersten Messung. Wie es zu einem ***Regressionseffekt*** kommt, läßt sich an einem vereinfachten Beispiel deutlich machen. Dazu nehmen wir an, in einem Stadtteil betrage ohne Einführung eines Förderungsprogramms der Durchschnitts-IQ der Kinder mit 4 Jahren 90 und derselben Kinder mit 5 Jahren wieder 90. Die bivariate Verteilung der IQ zu den beiden Meßzeitpunkten möge so aussehen, wie in Tabelle 9.1 angegeben.

Tabelle 9.1: Bivariate Häufigkeitsverteilung der IQ mit 4 Jahren und mit 5 Jahren in einem sozial benachteiligten Stadtteil (fingierte Daten)

		IQ mit 4 Jahren							
		75	80	85	90	95	100	105	Zeilensumme
	105	0	0	0	1	2	2	1	6
	100	0	0	2	15	28	15	2	62
IQ	95	0	2	28	90	90	28	2	240
mit 5	90	1	15	90	172	90	15	1	384
Jahren	85	2	28	90	90	28	2	0	240
	80	2	15	28	15	2	0	0	62
	75	1	2	2	1	0	0	0	6
Spaltensumme		6	62	240	384	240	62	6	

In Tabelle 9.1 sieht man, daß die Verteilung der IQ mit 4 Jahren und 5 Jahren gleich ist, zwischen den beiden Meßzeitpunkten besteht eine mittlere Korrelation von r = 0.5 (dieser Wert erscheint angesichts der Varianzeinschränkung als nicht unrealistisch).
Was hat man nun zu erwarten, wenn man unter den Vierjährigen alle Probanden auswählt, die einen IQ von 75 haben, und sie als Fünfjährige wieder untersucht? Greift man aus Tabelle 9.1 die entsprechende Zeile heraus und berechnet den Durchschnitt der IQ mit 5 Jahren, so findet man einen Mittelwert von 82.5. Geht man mit den anderen Zeilen aus Tabelle 9.1 entsprechend vor, so findet man, daß die Kinder, die mit 4 Jahren einen IQ von 80 hatten, im Durchschnitt mit 5 Jahren einen Wert von 85 haben, usw. (siehe Tabelle 9.2).

Tabelle 9.2: Regression der IQ mit 5 Jahren auf die IQ mit 4 Jahren, berechnet aus Tabelle 9.1

	Durchschnittlicher IQ
IQ mit 4 Jahren	mit 5 Jahren
105	97.7
100	95.0
95	92.5
90	90.0
85	87.5
80	85.0
75	82.5

Aus Tabelle 9.2 sieht man: Wenn man Kinder herausgreift, die mit 4 Jahren unter dem Mittelwert lagen, hat man mit 5 Jahren etwas höhere Werte zu erwarten, wohingegen man bei Kindern, die mit 4 Jahren über dem Mittelwert lagen, mit 5 Jahren im Durchschnitt etwas niedrigere Werte erhält. Dieser Effekt, den man ***Regressionseffekt*** nennt (eine ausführlichere Diskussion von Regressionseffekten findet man bei Stelzl, 1982, Kapitel 6), entspricht in etwas anderer Darstellung der Aussage, daß bei gleicher Verteilung der Vor- und Nachtestwerte die Nachtest-Vortest-Differenz (X_2 - X_1) negativ mit den Ausgangswerten korreliert. Der Regressionseffekt kann leicht mit Wirkungen des Förderungsprogramms verwechselt werden: Für das Programm werden gewöhnlich die Kinder mit besonders niedrigen Ausgangswerten als die Bedürftigsten ausgewählt. Eine solche Selektion läßt aber, wie gezeigt, auch ohne Behandlung ein Ansteigen der Werte erwarten. Der Erfolg der Fördermaßnahme muß also gegen diesen Regressionseffekt abgegrenzt werden.
Probleme bei der Zusammenstellung einer Kontrollgruppe: Aus den genannten Gründen, die zeigen, daß die Vortest-Nachtest-Differenz auch bei diesem Beispiel zur Schätzung des Programmeffekts wenig geeignet ist, erscheint es geboten, eine nicht behandelte Kontrollgruppe in den Versuchsplan mit einzubeziehen, um den Erfolg des Programms relativ zu dieser Kontrollgruppe beurteilen zu können (vgl. Beispiel 1). Da aber bei einem solchen Projekt neben den Zielsetzungen der Forschung vor allem soziale Gesichtspunkte zu berücksichtigen sind, können Versuchsgruppe und Kontrollgruppe nicht nach dem Zufall gebildet werden, wie es unter dem Gesichtspunkt der experimentellen Stringenz wünschenswert wäre. Man kann aber z.B. versuchen, in einem anderen Stadtteil Kinder zu finden, die den Projektkindern in den Ausgangswerten möglichst gut entsprechen, und diese Gruppe nach einem entsprechenden Zeitraum ebenfalls nachuntersuchen. Mit einer aus einem anderen Stadtteil zusammengestellten Kontrollgruppe sind allerdings systematische Unterschiede zur Projektgruppe in der Ausgangslage und be-

züglich der ohne Behandlung zu erwartenden weiteren Entwicklung nicht ganz auszuschließen: Eine Parallelisierung ist immer nur nach einer begrenzten Anzahl von Variablen möglich, und wenn sich die beiden Stadtteile hinsichtlich der sozialen Struktur stark unterscheiden, so ist zu erwarten, daß in einer ganzen Reihe von weiteren Variablen Restunterschiede zwischen Projektkindern und Kontrollkindern bestehen. Wenn ferner die Verteilung der IQ in den beiden Stadtteilen einen verschiedenen Mittelwert hat, so ist nach einer Selektion zum Zwecke der Parallelisierung mit unterschiedlichen Regressionseffekten zu rechnen (ein Argument, das im Zusammenhang mit der Evaluation von Programmen zur kornpensatorischen Erziehung vor allem von Campbell & Erlebacher (1975) ins Spiel gebracht wurde).

Trotz solcher Einwände ist jedoch keinesfalls zu übersehen, daß eine sorgfältig zusammengestellte Kontrollgruppe eine wesentliche Verbesserung des Versuchsplans darstellt: Effekte der Testwiederholung, von Verzerrungen in den Testnormen usw. sind kontrolliert. Regressionseffekte sind zwar nicht genau gleich gehalten, aber doch der Richtung und Größenordnung nach in etwa kontrolliert. Dasselbe gilt für allgemeine sozialpolitische Entwicklungen (sofern man nicht gerade das Pech hat, daß im fraglichen Zeitraum in nur einem der beiden Stadtteile ein besonderes sozialpolitisches Ereignis eintritt). Eine quasi-experimentelle Untersuchung mit einer gut gewählten Kontrollgruppe kann sicherlich genauso überzeugen wie eine experimentelle Prüfung - vorausgesetzt, die Effekte sind groß genug, daß man sich um ein oder zwei Punkte Unterschied durch ungleiche Regressionseffekte oder durch nicht perfekt kontrollierte Ausgangslage nicht zu streiten braucht.

Beispiel 3: Vergleich der Effektivität von Sonderschule und Regelschule bei leistungsschwachen Schülern (Probleme quasi-experimenteller Kontrolle)

Noch schwieriger als beim vorangehenden Beispiel, wo experimentelle Bedingungen doch zumindest näherungsweise realisiert werden können, ist die Situation dort, wo nur Feldforschung möglich ist. Wenn man z.B. untersuchen will, ob sich Kinder im IQ-Bereich von 80-90 an der Sonderschule oder an der Regelschule besser weiterentwickeln, so ist es natürlich ausgeschlossen, eine für die Betroffenen so schwerwiegende Entscheidung zu Forschungszwecken zu manipulieren. Damit bleibt nur die Möglichkeit, Paare von Kindern herauszusuchen, bei denen ähnliche Ausgangsbedingungen bestanden haben, wobei aber nur eines der beiden Kinder an die Sonderschule überwiesen wurde, während das andere an der Regelschule verblieb. Im folgenden sollen für diese Forschungssituation typische Probleme dargestellt werden. Auch hier soll deutlich werden, daß nicht unmittelbar augenfällige Fehlerquellen die Interpretation gefährden können, sofern man sie nicht zumindest der Größenordnung nach abschätzen und entsprechend in Rechnung stellen kann.

Probleme der Parallelisierung: Will man Paare von Kindern (jeweils ein Kind aus der Sonderschule, eines aus der Regelschule) zusammenstellen, bei denen gleiche Ausgangslage besteht, so wird man als Parallelisierungsmerkmale wohl in erster Linie das bis zum Zeitpunkt der Überweisung aufgetretene Ausmaß an Schulversagen und den IQ zu diesem Zeitpunkt heranziehen. Aber auch dann, wenn man eine Stichprobe von Sonderschülern und Regelschülern nach diesen Merkmalen parallelisiert hat und sie danach einige Jahre in ihrer weiteren Entwicklung beobachtet, kann man den Unterschied in der weiteren Entwicklung nicht ohne weiteres als Wirkung der Schule interpretieren. Die Parallelisierung ist zwar nach den vermutlich prognostisch wichtigsten Merkmalen (Schulversagen, IQ) erfolgt, trotzdem sind aber zwischen den Gruppen systemati-

sche Unterschiede zu erwarten: Wenn von zwei Kindern mit gleicher Intelligenz und gleichem Ausmaß an Schulversagen das eine an die Sonderschule, das andere an die Regelschule geschickt wird, so kommen als Gründe dafür nicht nur Zufälligkeiten des Entscheidungsprozesses und äußere Umstände in Betracht. Eine ganze Reihe von Entscheidungsgründen (Meinung des Lehrers; Einstellung der Eltern zur Schule, ihre Bereitschaft, Zusatzunterricht zu erteilen oder zu finanzieren; das Ausmaß, in dem das Kind in der bisherigen Klasse integriert ist, usw.) ist denkbar, und jeder dieser Gesichtspunkte kann für die weitere Entwicklung tatsächlich von großer prognostischer Relevanz sein. Vermutlich werden trotz der Parallelisierung die Ausgangsbedingungen und außerschulischen Umstände bei den Sonderschülern ungünstiger sein.

Probleme durch Selektionseffekte: Wenn man die Untersuchung nicht als Längsschnittstudie durchführt, was bekanntlich mühsam und langwierig ist, sondern retrospektiv, so hat man zusätzlich mit Selektionsproblemen zu kämpfen: Sucht man nämlich aus den jetzigen Regelschülern diejenigen heraus, die z.B. vor zwei Jahren genauso schlechte Leistungen hatten wie andere, die vor zwei Jahren an die Sonderschule überwiesen wurden, so kann man dabei offensichtlich nur diejenigen erfassen, die in den letzten beiden Jahren nicht so weit abgefallen sind, daß sie doch noch an die Sonderschule überwiesen worden wären. Diejenigen, die man jetzt immer noch an der Regelschule vorfindet, sind vermutlich bezüglich ihrer weiteren Entwicklung eine positive Selektion aus denen, die vor zwei Jahren trotz Schulversagens auf die Regelschule geschickt wurden. Bei den Sonderschülern hat keine entsprechende Selektion stattgefunden. Sofern man nicht zeigen kann, daß die Selektion im fraglichen Zeitraum vernachlässigenswert gering war, wird man damit rechnen müssen, daß die Studie einen Bias zu Ungunsten der Sonderschüler hat.

Schwierigkeiten der beschriebenen Art, insbesonders Kontrollgruppen mit ungleicher Ausgangslage, unvollständige Parallelisierung und Selektionseffekte sind für Feldstudien charakteristisch. Das sollte aber kein Grund sein, dort wo Evaluationsforschung nur mit Hilfe von Feldstudien betrieben werden kann, die Flinte ins Korn zu werfen. Man muß vielmehr versuchen, die Fehlerquellen zu erkennen und in ihrer Größenordnung abzuschätzen. Bei der oben diskutierten Fragestellung z.B. wirken sich alle genannten Fehlerquellen jeweils zuungunsten der Sonderschüler aus. Sollten als Ergebnis der Untersuchung die Sonderschüler besser abschneiden, so wird dieser Befund durch die genannten Störquellen nicht in Frage gestellt, sondern erscheint nur umso eindrucksvoller. Im Regelfall werden freilich Feldstudien zu Ergebnissen fuhren, deren Interpretation mit vielen Unsicherheiten belastet ist. In jedem Fall aber tragen sie dazu bei, die Diskussion um die Wirkung pädagogischer Maßnahmen auf eine empirische Grundlage zu stellen. In günstigen Fällen kann ihnen die gleiche Überzeugungskraft zukommen wie Experimenten, so daß es in den Hauptfragestellungen zu einer abschließenden Entscheidung kommt.

9.2.3 Braucht man zur Evaluation Forschung?

Bei Evaluationsforschung handelt es sich überwiegend um angewandte Forschung, bei der es darum geht, den Erfolg neu eingeführter Maßnahmen zu beurteilen. Das können soziale und pädagogische Maßnahmen sein, wie die erwähnten Programme zur kornpensatorischen Erziehung, aber auch medizinische Maßnahmen (z.B. eine Chemotherapie zusätzlich zur operativen Behandlung von Krebspatienten) oder auch

Änderungen der innerbetrieblichen Organisation (z.B. die Einführung gleitender Arbeitszeit). Im Vordergrund stehen Fragen der unmittelbaren Zielerreichung, der Vor- und Nachteile, Kosten und Nutzen der gesetzten Maßnahme.

Nun wird Evaluation und darauf aufbauende Programm-Modifikation auch laufend im Alltag betrieben, ohne daß dabei andere Methoden als die der Alltagserfahrung zum Einsatz kommen: Ein Lehrer, der aus den im Unterricht gegebenen Schülerantworten und den Lösungsversuchen bei Hausaufgaben und Klassenarbeiten Schlüsse auf den bislang erreichten Unterrichtserfolg zieht, um seinen weiteren Unterricht danach modifizierend zu gestalten, betreibt Evaluation ohne Einsatz wissenschaftlicher Methodik. Wenn in einem Projekt zur Nachbetreuung straffälliger Jugendlicher von den damit beauftragten Pädagogen ein jährlicher Rechenschaftsbericht gefordert wird, so erwartet der Auftraggeber darin primär eine Beschreibung des Ablaufs der von ihm finanzierten Maßnahme. Daß dabei positive Gesichtspunkte und Erfolge in den Vordergrund gestellt werden, wird er nicht anders erwarten, und nicht als Manipulation betrachten. Freilich wird er auch ein Minimum an "harten" Daten erwarten, z.B. Angaben über die Zahl der betreuten Personen, Häufigkeit der Kontakte, Zahl der im Betreuungszeitraum rückfällig Gewordenen, usw. Ein solcher Rechenschaftsbericht, der auch Vorschläge zur Programm-Modifikation enthalten kann, wird häufig ausreichen, um dem Auftraggeber eine Entscheidung über die Fortführung zu ermöglichen.

Wenn somit einerseits außer Zweifel steht, daß eine Beurteilung des Programmerfolgs und eine darauf aufbauende Programm-Modifikation auch mit Mitteln des alltäglichen Erfahrungslernens möglich ist (ähnlich wie Menschenbeurteilung auch ohne wissenschaftliche Diagnostik möglich ist), so ist andererseits nicht zu übersehen, daß dieser Weg zahlreiche Fehlerquellen enthält, denen man durch Einsatz einer wissenschaftlich kontrollierten Methodik entgegenwirken kann. Schon wenn es darum ginge, das Programm zur Nachbetreuung straffälliger Jugendlicher in größerem Umfang und mit erheblichem Kostenaufwand einzuführen, würde man das wohl kaum allein aufgrund eines Rechenschaftsberichts tun wollen, der die positiven Schilderungen der für das Projekt verantwortlichen Pädagogen enthält. Hier ist ein höheres Maß an Sicherheit der Aussagen erforderlich, so daß offensichtliche Fehlerquellen kontrolliert werden müssen, z.B. bewußte und unbewußte Verzerrungstendenzen der Berichterstatter. Man wird nicht nur nach dem Ablauf des Programms fragen, sondern an den Erfolg strengere Maßstäbe anlegen, z.B. einen Vergleich bezüglich verschiedener Kriterien der sozialen Eingliederung mit und ohne Betreuungsprogramm fordern. Auch wenn dann der Erfolg des Programms im vorliegenden Fall hinreichend sichergestellt ist, wird man nach der Verallgemeinerbarkeit fragen: Das Programm soll ja von anderen Pädagogen an anderen Orten übernommen werden. Was also sind die kritischen Bestandteile, die für den Erfolg ausschlaggebend waren (das Engagement der Pädagogen? der laufende Kontakt? spezielle Programmbestandteile, wie z.B. ein Training sozialer Fertigkeiten? bestimmte Merkmale der betreuten Probanden?) - diese Fragen sind aufgrund einer noch so gelungenen Projektschilderung nicht mit der Sicherheit zu beantworten, die nötig wäre, um eine großangelegte und teure Maßnahme einzuleiten.

Außer hohen Kosten können auch andere Grunde dafür sprechen, eine erhöhte Sicherheit der Aussagen zu fordern: Wenn die Maßnahme nur gegen Widerstand durchsetzbar ist (z.B. Koedukation) oder die Betroffenen erheblich belastet (Chemotherapie in der Medizin), so wird man sie nur dann durchführen wollen, wenn der Erfolg

bzw. die Wirksamkeit mit einiger Sicherheit nachgewiesen ist. In anderen Fällen ist der Einsatz von Forschung erforderlich, weil die Alltagserfahrung offensichtlich nicht ausreicht, um die anstehenden Fragen zu entscheiden: Etwa wenn aufgrund weltanschaulicher oder politischer Befrachtetheit der Thematik mit stark verzerrter Informationsverarbeitung zu rechnen ist, oder wenn sich aus anderen Gründen trotz eines langen Erfahrungszeitraums kein Konsens in der Beurteilung abzeichnet. In all diesen Fällen ist sozialwissenschaftliche Evaluationsforschung mit vollem Einsatz ihrer methodischen Möglichkeiten, insbesondere ihrer diagnostischen Instrumente und ihrer experimentellen und quasi-experimentellen Kontrolltechniken gefragt.

In den letzten Jahren vertreten nun einige Autoren, insbesonders in den USA (z.B. Stake, 1975; Guba & Lincoln, 1982), die Ansicht, daß als Alternative zur wissenschaftlichen ("scientific") Evaluation ein "anderes Paradigma" der Evaluationsforschung treten müsse, das sie "Responsive Evaluation" nennen. Letztere ist durch den Einsatz "naturalistischer" Methoden gekennzeichnet. Die bevorzugte Methode beim naturalistischen Ansatz ist die teilnehmende Beobachtung, die freie Beschreibung unter Betonung qualitativer Gesichtspunkte bei wechselnder Thematik. Das Programm, das es zu evaluieren gilt, liegt nicht fest, sondern kann während des Ablaufs modifiziert werden, u.a. auch auf Anraten des Evaluators als Experten ("invited interference"). Ein Bericht kann je nach Erfordernissen des Auftraggebers zu beliebigen Zeitpunkten gegeben werden. Er kann schriftlich oder mündlich erfolgen und orientiert sich am Informationsbedürfnis des Empfängers. Im Stil kann er sich eher an journalistischen Darstellungen als Vorbildern orientieren als an Experimentalberichten. - Ähnliche methodische Tendenzen, insbesonders was die Abkehr von festen Untersuchungsplänen zugunsten einer Mitgestaltung aller am Programm Beteiligten an Ablauf und Evaluation der Maßnahme anlangt, findet man im deutschen Sprachraum bei Vertretern der Aktions- und Handlungsforschung auf dem Hintergrund ihres speziellen wissenschaftstheoretischen Konzepts (Näheres siehe z.B. Kordes, 1984).

Was hier als naturalistische Methode beschrieben wird, entspricht indes über weite Strecken dem alltäglichen Erfahrungslernen, wobei allerdings bei der Person, die die Evaluation durchführt, besondere Qualitäten vorausgesetzt werden: Sie soll allen Informationsquellen gegenüber offen und sensitiv sein und an keinen Plan gebunden das jeweils Richtige tun bzw. dem Auftraggeber die geeigneten Maßnahmen empfehlen. Wie bereits im Zusammenhang mit der im Alltag ständig stattfindenden Evaluation ausgeführt, gibt es viele Fälle, in denen ein solches nicht regelgebundenes, "naturalistisches"Vorgehen und eine zwanglose Form der Berichterstattung ihren Zweck erfüllen: z.B. wenn es um die Lösung lokaler Probleme geht (etwa um eine Schulklasse, in der es besondere Spannungen gibt), oder auch wenn es um die Entwicklung und erste Erprobung von Programmen geht. Es wäre aber sicher naiv, anzunehmen, Sensibilität und guter Wille würden genügen, um Objektivität und Validität der Diagnostik zu gewährleisten und Behandlungseffekte von anderen Veränderungen (siehe oben: Unzulänglichkeiten des Vorher-Nachher-Versuchsplans) zu unterscheiden. So macht z.B. Wottawa (1981) eindrucksvoll deutlich, wie schwierig es ist, bei einem politisch umstrittenen Thema, wie dem Vergleich von Schulsystemen, manipulative Berichterstattung zu vermeiden. Nur der Einsatz hinsichtlich ihrer Gütekriterien überprüfter diagnostischer Instrumente und die weitestmögliche Offenlegung der Daten, die eine gegenseitige Kontrolle der Wissenschaftler ermöglicht, kann hier über den Meinungsstreit hinaus zu einem Erkenntnisfortschritt führen. "Naturalistischer"

Evaluation aber fehlt gerade das, was den spezifischen Beitrag der Wissenschaft ausmacht, nämlich die kontrollierte Methodik. Deshalb kann sie auch nicht, wie das Guba & Lincoln (1982) beanspruchen, im Sinn eines "Paradigmenwechsels" die "klassische" wissenschaftliche Evaluationsforschung ablösen.

Das kann sicher nicht heißen, daß nur Wissenschaftler die Welt verbessern könnten. Wenn es darum geht, in Problemsituationen Abhilfe zu schaffen, mögen andere Strategien (rasch wechselndes Ausprobieren, Propagieren von Lösungsvorschlägen auch bei niedriger Sicherheit, Beeinflussung von Entscheidungsträgern mit werbetechnischen Methoden usw.) erfolgreicher sein. Das sollte den Wissenschaftler aber nicht verleiten, den Wert, den seine Arbeit hat, und der auf Neutralität, methodische Kontrolle und Nachprüfbarkeit gegründet ist, gering zu achten und diese Qualitätsmerkmale, die in seiner Methodik liegen, aufzugeben oder auch herunterzuspielen, um statt dessen nach den Publicity-Erfolgen der Journalisten zu schielen.

Zusammenfassung

Ziel pädagogischer Evaluationsforschung ist es, wissenschaftlich fundierte Aussagen über die Wirkung von pädagogischen Maßnahmen zu machen. Dabei treten zu den Fragen der Auswahl der Erhebungsinstrumente (Tests, Beurteilungen, Interviews usw.) auch Fragen der Versuchsplanung und Auswertung. Solche Fragen wurden anhand von drei typischen Beispielen diskutiert.

Bei der Diskussion des ersten Beispiels (Verbalisieren beim Problemlösen) wurden folgende Probleme im Umgang mit Nachtest-Vortest-Differenzen behandelt: Das Problem des Skalenniveaus, das Reliabilitätsproblem, die negative Meßfehlerkorrelation zwischen Ausgangswerten und Zuwachs, das Problem der Abgrenzung des Behandlungseffekts gegen andere Veränderungen.

Probleme quasi-experimenteller Kontrolle wurden an Beispiel 2 und 3 illustriert. An Beispiel 2 (Frühförderung der kognitiven Entwicklung) wurde nochmals auf die Unzulänglichkeit eines einfachen Vorher-Nachher-Versuchsplans hingewiesen, hier insbesonders in Hinblick auf unkontrollierte Regressionseffekte. Selbst wenn Versuchsgruppe und Kontrollgruppe aus praktischen Gründen nicht nach dem Zufall gebildet werden können, stellt eine sorgfältig zusammengestellte Kontrollgruppe eine wesentliche Verbesserung des Versuchsplans dar. An Beispiel 3 (Vergleich der Effektivität von Sonderschule und Regelschule bei leistungsschwachen Schülern) werden Probleme dargestellt, die auftreten, wenn durch Selektion aus unterschiedlichen Populationen (Sonderschüler versus Regelschüler) zwei Gruppen mit gleicher Ausgangslage zusammengestellt werden sollen.

Bei der Vielfalt möglicher Fragestellungen, die im Rahmen pädagogischer Evaluationsforschung auftreten können, konnte es nicht das Ziel dieses Kapitels sein, einen repräsentativen Überblick über die Methodik zu geben. An den dargestellten Beispielen sollte aber doch deutlich geworden sein, daß Nachweis und Analyse von Programmeffekten einer sophistizierten Methodik bedürfen, die nicht durch teilnehmende Beobachtung und freie Beschreibung des Programms zu ersetzen ist.

Einführende Literatur:

Wottawa, H. & Thierau, H. (1989). ***Evaluation.*** Bern: Huber.

Weiterführende Literatur:

Cronbach, L.J.(1983). ***Designing evaluations of educational and social programs*** (2nd ed.). San Francisco: Jossey-Bass Publishers.

Krauth, J. (1983). Methodische Probleme in der pädagogischen Evaluationsforschung. ***Zeitschrift für Empirische Pädagogik, 7,*** 1 -21.

Wittmann, W.W. (1985). ***Evaluationsforschung. Aufgaben, Probleme und Anwendungen.*** Berlin: Springer.

Weitere Diskussionsbeiträge und inhaltliche Beispiele

findet man in der Zeitschrift für Pädagogische Psychologie, 4, 1990, Heft 4 (Themenheft zur Evaluationsforschung).

Zur Auseinandersetzung mit der Aktionsforschung:

Patton, M.Q. (1981). ***Creative Evaluation.*** London: Sage.

Kordes, H. (1984). Aktionsforschung. In H. Haft & H. Kordes (Hrsg.), ***Methoden der Erziehungs- und Bildungsforschung.*** Enzyklopädie der Erziehungswissenschaften, Bd. 2 (S. 185-219). Stuttgart: Klett-Kotta.

Zecha, G. & Lukesch, H. (1982). Die Methodologie der Aktionsforschung. Analyse, Kritik, Konsequenzen. In J.L. Patry (Hrsg.), ***Feldforschung (S.367 - 387).*** Bern: Huber.

10. Pädagogische und psychologische Aspekte

1. Was versteht man in der Pädagogisch-psychologischen Diagnostik unter Schulleistung, und wie kommt Schulleistung zustande?
2. Welche diagnostischen Parameter bestimmen das Ergebnis von Leistungsmessungen?
3. Wann und wie häufig soll diagnostiziert werden?
4. Welche Nebenwirkungen und Fehler können auftreten?

Vorstrukturierende Lesehilfe

In der diagnostischen Praxis geht es um die Anwendung von Regeln und Methoden, die es gestatten, individuelle Merkmalszustände vor, während und nach einer pädagogischen Behandlung so genau wie nötig zu erfassen, damit die darauf gestützten Entscheidungen zum bestmöglichen Erfolg führen (Optimierungsgrundsatz).

Dem besseren Verständnis des Sachzusammenhangs dient ein Exkurs, in dem pädagogische Fachausdrücke (wie Didaktik, Curriculum und Lehrziel) erläutert werden. Als zentraler Sachverhalt wird das Konstrukt "Schulleistung" eingeführt. Es ist von Schüler- wie von Schulmerkmalen determiniert und kann über Indikatorvariablen (Lehrerurteile, Testwerte) erfaßt werden.

Zur Messung schulischer Lernfortschritte bedarf es der validen Operationalisierung der Lehrziele durch den Unterricht und einer dafür repräsentativen Auswahl von Aufgaben. Die Ergebnisse werden außerdem - und neben den Schülerparametern - vom Meßzeitpunkt beeinflußt. Der didaktische Nutzen diagnostischer Maßnahmen hängt u.a. von der Häufigkeit der Messung (Meßdichte) ab; zwischen beiden wird eine kurvilineare Beziehung angenommen.

Die Messung von Schulleistungen kann u.U. in pädagogisch unerwünschter Weise auf den Unterricht zurückwirken und sozialpsychologische Belastungen mit sich bringen. Ebenso zu beachten sind mögliche Erwartungseffekte sowie Attribuierungs- und Beurteilungsfehler, die die Ergebnisse von Schätzverfahren verzerren können.

10.1 Die Funktion Pädagogisch-psychologischer Diagnostik

Die Pädagogisch-psychologische Diagnostik richtet sich wie jede andere Diagnostik auf die Erhebung individueller Ist-Zustände, d.h. diagnostische Aussagen enthalten Informationen über die Ausprägung interessierender Merkmale bei einzelnen Merkmalsträgem. Auch Aussagen über Gruppen beruhen darauf. Sie ordnen Individuen in bezug auf das betrachtete Merkmal homogenen Klassen von Merkmalsträgem zu (vgl.

z.B. Rollett, 1978; allgemein Kallus & Janke, 1988). Die Anzahl und die Breite der zu bildenden Klassen ergeben sich aus der Bedeutung der Klassifikation für die pädagogischen Zielvorgaben und Behandlungszuweisungen sowie der zu erwartenden Folgen. Dies entspricht dem zentralen lernorganisatorischen Problem der Bildung hinreichend homogener bzw. heterogener Lerngruppen, gleich ob es sich um Ein- oder Umschulungsentscheidungen, Kurszuweisungen, die ad-hoc-Bildung von Lerngruppen im Gruppenunterricht oder um die Behandlung von Lernstörungen handelt.

Aus der Unterscheidung von Ist- und Soll-Zuständen ergibt sich eine zweifache Funktion für die Pädagogisch-psychologische Diagnostik, nämlich zum einen die Feststellung von Lernvoraussetzungen vor einer geplanten oder erwogenen Intervention (Lernsteuerung), zum anderen die Feststellung des Lern- oder Behandlungserfolgs nach einer Intervention (Lernkontrolle; Diagnostik als Rückkoppelungsglied; Pawlik, 1982, S. 22). Unabhängig von den pädagogischen Absichten und den Randbedingungen ist beides, diagnostisch gesehen, dasselbe: die Beschreibung individueller Ist-Zustände zu verschiedenen Zeitpunkten. Soweit es sich um wiederholte Messung desselben Merkmals (an denselben Personen) handelt, können individuelle Veränderungen diagnostiziert werden. Deren Bedeutung läßt sich jedoch nur einschätzen, wenn man dabei auf Unterschiede zwischen Individuen zurückgreift (vgl. Abschnitt 9).

Diagnostische Feststellungen - seien es die alltäglichen informellen, seien es formalisierte - werden in der Regel im Hinblick auf pädagogische Entscheidungen getroffen, d.h. im Hinblick auf bestimmte, für nötig und/oder möglich gehaltene Verhaltensänderungen. Sie sind also mit Erwartungen verknüpft: Neben den retrospektiven wird ihnen eine prospektive Bedeutung beigemessen, die den Charakter einer Vorhersage hat (Wenn-Dann-Verknüpfung: $\rightarrow$); z.B. stützen wir die Entscheidung, ein Kind an die Schule für Lernbehinderte zu überweisen, auf die Erwartung, daß es dort besser gefördert werden kann. Pädagogisches Handeln findet in der Regel vor dem Hintergrund komplexer Erwartungsgeflechte statt, wobei vielfach die Erfolgswahrscheinlichkeiten von Handlungsalternativen gegeneinander abzuwägen sind (vgl. Westmeyer, 1978; allgemein Noack & Petermann, 1988).

Zumindest für formalisierte Verfahren und für Entscheidungen von erheblicher Tragweite gilt daher, daß die diagnostischen Informationen hohen Ansprüchen an ihre prädiktive Validität genügen müssen. Die wesentliche Funktion der pädagogisch-psychologischen Diagnostik besteht darin, "richtige" Entscheidungen herbeizuführen und damit die Wahrscheinlichkeit des Erfolgs der pädagogischen Behandlung zu erhöhen. Es liegt auf der Hand, daß Veränderungen in der Regel umso besser bewirkt werden können, je genauer und valider der Ist-Zustand erhoben wird und je genauer der Soll-Zustand definiert ist. Pädagogische Praxis ohne Diagnostik wäre blind; Diagnostik ohne Praxisbezug wäre bedeutungsleer (Tent & Waldow, 1984).

Die diagnostische Praxis besteht aus der Anwendung eines fundierten technologischen Regelwissens. Ihre Aufgabe besteht zusammenfassend darin, zur Optimierung pädagogischen Handelns beizutragen, indem sie (idealtypisch)

(a) die tatsächlichen Ausgangsbedingungen bei den Lernenden klärt (Diagnose von Lernvoraussetzungen)
(b) die Wahrscheinlichkeit der Folgen einschätzt, die unter gegebenen Bedingungen bei dieser oder jener Behandlungsalternative zu erwarten sind *(auf Grundlagenforschung gestützte Prognose)*
(c) die tatsächlichen Folgezustände bei den Lernenden feststellt (Diagnose des Lern- oder Behandlungserfolgs).

10.2 Didaktischer Exkurs

Wie jedes andere ist pädagogisches Handeln an Zielen orientiert. Läßt man die in der Erziehungswissenschaft verbreitete Unterscheidung zwischen Erziehung, Bildung, Ausbildung, Lehre und Unterricht beiseite, kann man mit Blick auf die Schule summarisch von Lehrzielen sprechen. Diese im einzelnen zu formulieren, ist Aufgabe der Didaktik. Didaktik kann als das Kerngebiet der Erziehungswissenschaft verstanden werden. Ihr Gegenstandsbereich ist die Theorie des Unterrichts. Sie umfaßt alle Aspekte der Ziele, der Inhalte und der Organisation von Unterricht, unter Einschluß der Begründungen und der Voraussetzungen. Die allgemeine Didaktik wird in den Fachdidaktiken für die einzelnen Unterrichtsfächer und -gebiete sowie in den Stufendidaktiken für einzelne Schulstufen spezialisiert. Die Erkenntnisse der Didaktik finden ihren Niederschlag in den Lehrplänen, Richtlinien und Curricula, die den Schulen in der Regel von staatlichen Instanzen vorgegeben werden. Vereinfacht gesagt, geben Lehrpläne, Richtlinien und Curricula an, was, wann, weshalb, wozu und wie unterrichtet werden soll. Sie lassen dabei dem Lehrer mehr oder weniger große Spielräume, wie er die Vorgaben umsetzt, z.B. welche Leselehrmethode er benutzt, welches von mehreren zugelassenen Schulbüchern er seinem Lateinkurs zugrundelegt oder an welchem Drama er "exemplarisch" das Absurde Theater behandelt. Zwischen Lehrplan und Curriculum wird zumeist in der Weise unterschieden, daß ein Curriculum neben der Auflistung von Zielen, Inhalten und Unterrichtsmethoden auch die von der Didaktik erarbeiteten Begründungszusammenhänge darlegt und Reformansprüche einlösen will. Lehrpläne und Curricula stellen die verbindlichen Grundlagen für die systematische Planung, Durchführung und Auswertung von Unterrichtssequenzen dar.

Die Erziehungs- und Lehrziele beschreiben Soll-Zustände (Normen), auf die hin Merkmale von Lernenden verändert werden sollen. Pädagogische Absichten und Handlungen sind darauf gerichtet, das tatsächliche Verhalten der Lernenden den Soll-Zuständen möglichst weitgehend anzunähern. Den Lehrzielen können auf seiten der Lernenden analoge Lernziele zugeordnet sein. Lernziele sind subjektive Normvorgaben für die Änderung eigener Personmerkmale. Gelegentlich wird mit Lernziel auch das nach erfolgreicher Behandlung erworbene Verhalten bezeichnet.

Lehrziele kommen in der Praxis natürlich nicht isoliert vor. Sie sind in der Regel in bestimmter Weise miteinander verknüpft. Innerhalb eines Schulfachs bauen sie vielfach aufeinander auf, und sie ermöglichen oder erleichtern die Verwirklichung der Lehrziele in anderen Fächern. Sie bilden insgesamt eine nach sachlogischen und didaktischen Gesichtspunkten geordnete Struktur. Die Position der einzelnen Teilziele ist nicht beliebig vertauschbar. Damit ist zugleich ihre zeitliche Aufeinanderfolge mehr oder weniger eindeutig festgelegt (vgl. z.B. Klauer, 1974; Möller, 1974, 1976; Schott, Neeb & Wieberg, 1981). Aus diagnostischer wie auch aus didaktischer Sicht stellt sich die Frage, wieweit es möglich und zweckmäßig ist, Teil-Lehrziele aus dem Lehrzielverbund herauszulösen und diagnostisch zu isolieren; allgemein gefragt, wie groß die Lehrzielausschnitte bzw. die diagnostischen Einheiten sein und wo innerhalb einer Lehrzielanordnung sie liegen sollen. Für die Zwecke der Pädagogisch-psychologischen Diagnostik ist ein Lehrziel operational durch eine repräsentative Aufgabenstichprobe zu definieren.

Die Vielfalt der Lehrziele läßt sich nach verschiedenen Klassifikationsgesichtspunkten ordnen. Man spricht von einer Lehrzielhierarchie oder Lehrzieltaxonomie,

wenn Lehrziele konsistent nach einem formalen theoretischen Kriterium, z.B. der Komplexität oder dem Abstraktionsniveau, gereiht werden. Mehrdimensionale Anordnungen werden als Lehrzielmatrix bezeichnet. Ein solches Ordnungsgerüst liegt z.B. vor, wenn inhaltlich definierte Teilbereiche eines Unterrichtsfachs wie Gesellschaftslehre mit Verhaltensklassen oder den psychologischen Kategorien kognitiv, affektiv und motivational kombiniert werden. Zwei unterschiedliche Beispiele für Lehrzielmatrizen sind im Kasten S. 209-211 wiedergegeben; vgl. Abschnitt 6.2.3.

Ungeachtet der Wertungen, die darin zum Ausdruck kommen, sind Lehrzieltaxonomien und Lehrzielmatrizen unter dem Blickwinkel der Diagnostik deskriptive Ordnungsschemata für geforderte Schulleistungen. Neben dem fachlich-inhaltlichen Aspekten sind dabei vor allem der Abstraktionsgrad der Lehrziele und die Eindeutigkeit von Interesse, mit der sie spezifiziert werden, weil die diagnostischen Verfahren möglichst paßgenau darauf abgestimmt sein müssen (curriculare Validität). Hier ist eine Abstufung etwa im Sinne von Möller (1976) hilfreich, die zwischen Richtzielen, Grobzielen und Feinzielen unterscheidet.

Als Richtziele werden Lehrziele des Abstraktionsniveaus 3 bezeichnet. Sie weisen den geringsten Grad an Eindeutigkeit und Präzision auf. Sie werden in umfassenden, wenig spezifischen Begriffen formuliert, die die Richtung deutlich machen, in der gelernt werden soll, ohne daß damit bereits ein bestimmtes Verhalten festgelegt wird. Die Verwirklichung solcher Lehrziele kann sich also in einer Vielzahl z.T. sehr verschiedener konkreter Verhaltensweisen äußern Ziele dieser Art finden sich als Leitvorstellungen in Verfassungsartikeln, Schulgesetzen und Einleitungen zu Curricula. So schreibt z.B. die Verfassung des Landes Nordrhein-Westfalen von 1950 in Artikel 7 vor: "[Erziehungsziel] (1) Ehrfurcht vor Gott, Achtung vor der Würde des Menschen und Bereitschaft zum sozialen Handeln zu wecken, ist vornehmstes Ziel der Erziehung. (2) Die Jugend soll erzogen werden im Geiste der Menschlichkeit, der Demokratie und der Freiheit, zur Duldsamkeit und zur Achtung vor der Überzeugung des anderen, in Liebe zu Volk und Heimat, zur Völkergemeinschaft und Friedensgesinnung."

Grobziele sind Lehrziele vom Abstraktionsniveau 2 mit mittlerer Eindeutigkeit und Präzision. Es werden bereits Inhalte angegeben, in denen sich die allgemein gehaltenen Richtziele manifestieren. Die Beschreibung läßt aber noch verschiedene Varianten des Endverhaltens zu, und es fehlt an einem eindeutigen Maßstab für die Beurteilung des Erfolgs; z.B. in der Geometrie die Kongruenzsätze oder in der lateinischen Syntax den ablativus absolutus kennen.

Demgegenüber sind die Feinlehrziele voll operationalisiert (Abstraktionsniveau 1). Sie beschreiben das erwünschte Endverhalten genau und enthalten den Maßstab zu dessen Beurteilung; z.B. die Länge der Diagonale in einem Quadrat bestimmen oder ein Kapitel aus Caesars "Gallischem Krieg" übersetzen können. Darüber hinaus wird von Mikrolehrzielen (oder Feinstlehrzielen) gesprochen, wenn es um die konkreten Teillernschritte einer Lehrsequenz innerhalb einer Unterrichtsstunde geht.

Diagnostisch bedeutungsvoll ist auch das Verhältnis der Lehrziele zu den Lernzielen. Man kann nicht ohne weiteres erwarten, daß sie sich von vornherein entsprächen oder sogar deckten. Unter den Bedingungen der Schulpflicht und des "Massen-" Unterrichts bedarf es häufig besonderer didaktischer Bemühungen, eine hinreichende Korrespondenz herbeizuführen. Sie liegt in dem Maße vor, wie der innere Zustand der Lernenden, d.h. die Richtung und Intensität ihrer motivationalen Bereitschaft, die intendierte Beeinflussung zuläßt. Diese notwendige Bedingung kann unterschiedlich

Zur Klassifikation von Lehrzielen

Sachgerechter Unterricht erfordert - neben anderen Voraussetzungen - eine genaue Lehrstoffanalyse, die Präzisierung der Lehrziele, die Auswahl angemessener Lehrmethoden und Lehrmittel sowie die Überprüfung des Lehrerfolgs. Ordnungsschemata von der Art der ***Lehrzieltaxonomien*** und ***Lehrzielmatrizen*** sind als Raster zu verstehen, die bei der Planung von Unterrichtssequenzen hilfreich sein können. Sie spezifieren die Inhalte, die im Unterricht vermittelt werden sollen (Aneignungsphase) und dementsprechend in die Erfolgskontrolle einzubeziehen sind (Prüfungsphase). Darüber hinaus verdeutlichen sie die Beziehungen zwischen den einzelnen Teilzielen, insbesondere deren hierarchische Abhängigkeit voneinander und legen insoweit auch die zeitliche Abfolge der Lehrschritte fest. Als idealtypisch vereinfachte Strukturierungshilfen können sie das, was in dem hochkomplexen Bedingungsgefüge "Unterricht" tatsächlich geschieht, allerdings nur näherungsweise vorbestimmen.

(1) Eine Lehrzielmatrix zur "Fähigkeit eines Kindes, Konflikte zu lösen" (nach Schott, Neeb & Wieberg, 1981, S. 25).

				Handlungsaspekt					
				Soziale Verhaltensweisen					
				eigene Bedürfnisse äußern	eigene Bedürfnisse verteidigen	auf eigene Wünsche verzichten	Bedürfnisse anderer erkennen	Bedürfnisse anderer akzeptieren	Kompromisse schließen
				A	B	C	D	E	F
Situativer Aspekt	Arten von Kommunikationspartnern	vertraute Erwachsene	Einzelperson 1						
			Gruppe 2						
		unvertraute Erwachsene	Einzelperson 3						
			Gruppe 4						
		vertraute Gleichaltrige	Einzelperson 5						
			Gruppe 6						
		unvertraute Gleichaltrige	Einzelperson 7						
			Gruppe 8						

Diese Matrix setzt verschiedene ***situative Bedingungen*** mit spezifischen ***Handlungskompetenzen*** in Beziehung, die das Kind erwerben soll. Die einzelnen Zellen repräsentieren Teillehrziele, die nach Maßgabe der relevanten Randbedingungen ausgewählt werden können. Das Feld E8 z.B. beschreibt das Lehrziel, in einer unvertrauten Gruppe Gleichaltriger die handlungsbestimmenden Bedürfnisse der anderen zu akzeptieren. Dies setzt das Erkennen der Bedürfnisse (D8) voraus. Zu den Randbedingungen zählen der (kognitive) Entwicklungsstand des Kindes, die situativen Anforderungen, die in seiner Lernumwelt vorkommen sowie seine Fähigkeit zum Lerntransfer.

(2) Eine Matrix zur Klassifikation von Lehrzielen für die Unterrichtseinheit "Entwicklungsland Nigeria" im Erdkundeunterricht des 8. oder 9. Schuljahrs (nach S. Schacht, in Möller, 1974, S. 143-152). Die gekennzeichneten Felder werden im Textdurch Beispiele belegt.

Inhaltsklassen	Verhaltensklassen					
	Wissen (1.00)	Verstehen (2.00)	Anwenden (3.00)	Analyse (4.00)	Synthese (5.00)	Beurteilung (6.00)
I. Naturlandschaftsgeographie (Physische Geographie, Geo-Ökologie — ohne Unterklassen)						
II. Kulturlandschaftsgeographie (Anthropogeographie)						
II.1 Bevölkerungsgeographie						
II.2 Siedlungsgeographie						
II.3 Agrargeographie						
II.4 Industriegeographie	+	+	+	+	+	+
II.5 Geographie des tertiären Sektors (Handels-, Verkehrsgeographie)						
II.6 Politische Geographie						
III. Sozialgeographische Betrachtungsweisen						
III.1 Sozialräumliche Ordnungen						
III.2 Regionale Systeme						
III.3 Sozialgeographische Prozesse (z. B. Innovationen)						
III.4 Prognosen						
IV. Methodische Hilfen						
IV.1 Karten						
IV.2 Terrestrische Bilder, Luftbilder						
IV.3 Quantitative Methoden und Modelle						

Die Inhaltsklassen 1 bis IV orientieren sich an einem bestimmten geographischen Ordnungsschema, die Verhaltensklassen 1.00 bis 6.00 an der "kognitiven Lernzieltaxonomie" von Bloom und Mitarbeitern (1971). Die Ziffern besagen hier, daß die Taxonomieklassen nach Bedarf bis zu zwei Stellen weiter unterteilt werden können. Die Zellen beschreiben mög***liche Feinlehrziele:*** die vom Lehrer ausgewählten Feinlehrziele sind bestimmten ***Grobzielen*** zugeordnet und müssen in der Matrix lokalisierbar sein. Die Unterrichtseinheit umfaßt in diesem Fall elf Grobziele, u.a.: "Der Schüler soll wichtige Daten und Fakten über die Bevölkerung Nigerias und ihre tribalistische Differenzierung kennen" (Grobziel [GZ] 1, zu Inhaltsklasse [IK] II.l), "Der Schüler soll Entwicklungsplanungen und Investitionen am Beispiel der Region Ostnigeria (Biafra) beurteilen, um die Problematik von Entwicklungshilfe zu durchschauen" (GZ 7, zu IK II.4), "Der Schüler soll die Entwicklung, Bedeutung und gegenwärtige Situation der nigerianischen Erdölwirtschaft kennen" (GZ 8, zu IK II.4). Die sechs Taxonomieklassen werden hier durch die folgenden Beispiele für Feinlehrziele zur Inhaltsklasse II.4 erläutert:

Klasse 1.00: Wissen (zu GZ 8)
Der Schüler soll aus einer Liste mit vier Eigenschaften diejenige heraussuchen und ankreuzen, die das nigerianische Erdöl besonders wertvoll macht.

Klasse 2.00: Verstehen (zu GZ 8)
Der Schüler soll auf Grund einer vorgegebenen Tabelle über die Entwicklung der Erdölproduktion vom Jahre 1958 bis zum Jahre 1970 in ein vorgegebenes Koordinatensystem ein entsprechendes Kurvendiagramm einzeichnen.

Klasse 3.00: Anwenden (zu GZ 8)
Der Schüler soll durch Aufschreiben weniger Stichworte den Tiefpunkt der Erdölproduktion im Jahre 1968 begründen (Bürgerkrieg Nigeria - Biafra; Seeblockade des Ölhafens Bonny).

Klasse 4.00: Analyse (zu GZ 7)
Der Schüler soll aus zwei ihm vorgegebenen Informationen (1. Europäische und amerikanische Erdölgesellschaften haben in Nigeria mehrere Mrd. Dollar investiert, 2. Die Sowjetunion will sich mit 55 Mill. £N am Aufbau eines Stahlwerkes in Nigeria beteiligen) die Interessen der Großmächte in Nigeria ableiten. Er soll aufschreiben und schriftlich diskutieren, inwiefern es hier zu einem doppelten Interessenkonflikt kommt (Ost -West; Industrienationen - Entwicklungsländer).

Klasse 5.00: Synthese (zu GZ 7)
Der Schüler soll an Hand einer vorgegebenen Karte von Ostnigeria genau zwei optimale Standorte für ein geplantes Stahlwerk benennen und wenigstens fünf Gründe aufschreiben, die jeweils für die Wahl eines Standortes sprechen.

Klasse 6.00: Beurteilung ("Evaluation"; zu GZ 7)
Der Schüler soll einen vorgegebenen Schulbuchtext über die nigerianische Erdölwirtschaft (Dreimal um die Erde, Bd. 2, S. 21) kritisch beurteilen. Er soll erkennen und sinngemäß aufschreiben, daß in dem Text zwar auf die hohen Investitionen und Abgaben hingewiesen wird, daß aber über Eigentumsverhältnisse und Gewinne kein Wort fällt.

In den Feinlehrzielen sind teilweise zugleich die Kriterien für den Lernerfolg festgelegt.

gut erfüllt sein. Sie geht indirekt in die Varianz vieler Merkmale ein, die uns diagnostisch interessieren, ohne daß wir jeweils ihren Anteil erkennen könnten. So schwer es oft ist, dies nachzuweisen (vgl. Tent, 1969, S. 135-140; Kühn, 1983, S. 159-163), wir müssen davon ausgehen, daß motivationale Komponenten Bestandteil der Meßergebnisse bei allen Leistungen sind, die der pädagogischen Beeinflussung unterliegen.

Häufig wird es notwendig sein, die motivationalen Voraussetzungen im Einzelfall zu klaren.

Grundlegende Literatur zu 10.2:

Brezinka, W. (1981). Erziehungsziele, Erziehungsmittel, Erziehungserfolg. Beiträge zu einem System der Erziehungswissenschaft (2. Aufl.) München: Reinhardt.
Glöckel, H. (1992). Vom Unterricht. Lehrbuch der Allgemeinen Didaktik (2. Aufl.). Bad Heilbrunn: Klinkhardt.
Roth, L. (Hrsg.) (1991). Pädagogik. Handbuch ***für*** Studium und Praxis. München: Ehrenwirth.

Weiterführende Literatur:

Diederich, J. (1988). Didaktisches Denken. Weinheim: Juventa.
Klafki, W. (1991). Neue Studien zur Bildungstheorie und Didaktik (2. Aufl.). Weinheim: Beltz.

10.3 Schulleistung als Konstrukt

Konzentriert man sich auf die Schule, ist der zentrale Gegenstand Pädagogisch-psychologischer Diagnostik das, was man summarisch als "Schulleistung" bezeichnet. Es werden Schülermerkmale erfaßt, für die entweder evident ist, daß es sich um Schulleistungen oder Verhaltensaspekte im Sinne der üblichen Zeugnisrubriken handelt (z.B. Aufsätze verfassen, Prozentrechnung; Betragen, Fleiß), oder es geht um Merkmale, die systematisch damit korrelieren und sich anteilig als Voraussetzung oder als Folge pädagogischer Behandlung interpretieren lassen (z.B. Intelligenz, Konzentration, Ängstlichkeit). Diagnostische Aussagen können sich dementsprechend auf alle individuell bedeutungsvollen Aspekte von Schulleistung erstrecken. Dies schließt neben kognitiven und motorischen auch solche Schülermerkmale ein, die sozialen, affektiven und motivationalen Lehrzielen zugeordnet sind, wie Einstellungen, Werthaltungen und Gesinnungen. Mit "Leistung" wird hier im allgemeinsten und wertungsneutralen Sinn jedes Ergebnis menschlichen Handelns bezeichnet.

Schulleistungen liegen nicht einfach offen zutage. Sie basieren zwar auf Beobachtungsdaten; deren Bedeutung muß jedoch jeweils erst ermittelt werden. Testwerte, Schulnoten und andere Lehrerurteile haben dabei die Funktion von Indikatorvariablen. Sie stellen unterschiedliche operationale Definitionen dar, in denen sich Schulleistungen mehr oder weniger konkordant manifestieren. Als allgemeine schulpädagogische Kategorie wird "Schulleistung" trotz des Grundwortes -leistung zum Sammelbegriff für eine Vielzahl pädagogisch-psychologischer Einzelkonstrukte, die als Resultanten eines multifaktoriellen, heterogenen Beziehungsgeflechts aus Schüler- und Schulmerkmalen zu verstehen sind (Tent, Fingerhut & Langfeldt, 1976, S. 15-18; Langfeldt & Fingerhut, 1984; s. Abbildung 10.1, S. 213).

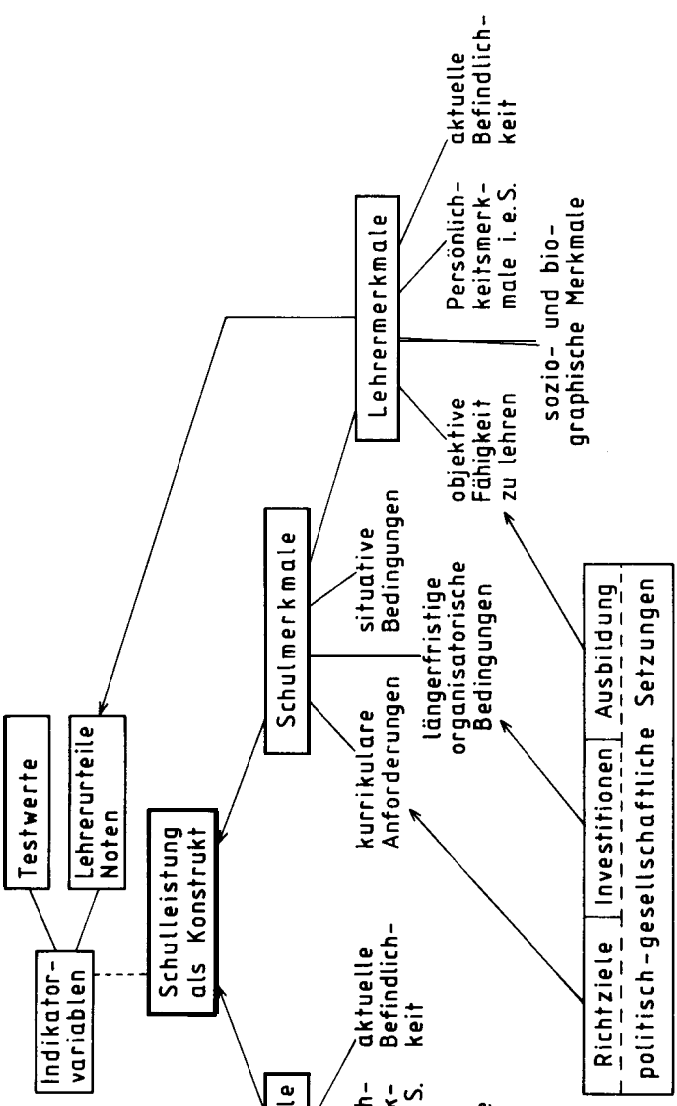


Abbildung 10.1: Das Konstrukt Schulleistung. Vereinfachtes Schema seiner Determinanten und Indikatoren.

Zu den "Schulmerkmalen" zählen vor allem die Lehrer, aber auch die Lehrpläne und die organisatorischen Bedingungen, wie z.B. die Gliederung des Schulsystems, die Lehrer-Schüler-Relation (Meßzahl) und die materielle Ausstattung der Schulen. Die Beschaffenheit der Schulmerkmale geht im wesentlichen auf gesellschaftliche Übereinkünfte zurück, die von Legislative und Exekutive als den politisch verantwortlichen Entscheidungsträgern in verbindliche Vorschriften (Gesetze, Verordnungen, Erlasse) umgesetzt werden.

Die einzelnen Wirkgrößen tragen nicht nur für sich zum Zustandekommen von Schulleistung bei. Nach den in Abschnitt 1.4.1 dargelegten Modellannahmen ist mit Kovarianz- und Wechselwirkungseffekten sowohl innerhalb der Schüler- wie der Schulmerkmale als auch zwischen beiden zu rechnen. Bemerkenswerte Beispiele für Wechselwirkungen zwischen Unterrichtsstil und Schülertyp hat Bennett (1979) für Leistungen in Mathematik und Muttersprache gefunden; Rheinberg (1980) konnte für die Furchtreduktion bei Schülern eine Wechselwirkung zwischen der Intelligenz der Schüler und der Bezugsnormorientierung bei der Leistungsrückmeldung durch die Lehrer nachweisen.

Die zwei Klassen von Indikatorvariablen für das Konstrukt unterscheiden sich in der Regel durch ihre instrumentelle Güte. Als Schätzwerte für die Resultante "Schulleistung" verkörpern Schulnoten und andere Lehrerurteile die diagnostische Leistung eines Beteiligten; d.h. sie sind zugleich anteilig Selbsteinschätzung. Wie gut sie im konkreten Fall sind, ist zumeist nicht bekannt. Demnach können empirische Schulnoten die individuellen Unterschiede bei den Schülern von vornherein nur unvollkommen widerspiegeln, weil kaum vermeidbare konstante oder variable Lehreranteile unkontrolliert in die Urteilsbildung einfließen. Dies schließt nicht aus, daß Lehrerurteile hoch valide sein können, aber man kann es ihnen, anders als bei einem standardisierten Test mit bekannten Güteeigenschaften, nicht "ansehen". Allerdings gilt - wie schon früher betont - auch für objektive Schulleistungstests, daß sie über die Schülerleistung indirekt die Qualität des erteilten Unterrichts miterfassen. Schulleistungen werden zwar an Schülern erhoben und ihnen auch diagnostisch zugeschrieben; die Abbildung 10.1 macht jedoch deutlich, daß diese einseitige Kausalattribuierung unzulässig ist, weil die übrigen Bedingungen nicht annähernd konstant gehalten werden können und ihr Varianzanteil zumeist nicht bekannt ist.

Pädagogisch-psychologische Diagnostik richtet sich sowohl auf die Schulleistung als Resultante als auch auf die übrigen Personmerkmale der Beteiligten. Bei den Schülern sind es Merkmale, die hier zusammenfassend als individuelle Lernvoraussetzungen bezeichnet werden können. Grundsätzlich kann sich Diagnostik aber auch auf die analogen Merkmale bei den Lehrern erstrecken. Anders als in den USA (s. Millman & Darling-Hammond, 1990) wird dieser Anteil des Bedingungsgefüges bei uns relativ wenig beachtet (zur Methodik vgl. Bessoth, 1983). Zur didaktischen Kompetenz der Lehrer gehört die Fähigkeit, Verhalten und Leistungen der Schüler zu beurteilen. Der Erfolg pädagogischen Handelns kann nicht besser sein als die Diagnose der Ausgangsbedingungen, an denen es ansetzt. Die diagnostische Kompetenz der Lehrer spiegelt sich daher nicht nur in der Qualität der Notengebung, sie schlägt sich auch im objektiven Schulerfolg nieder (Schrader, 1989). Die übrigen im Schema angeführten nichtpersonalen Merkmale und Wirkfaktoren haben, auf die Diagnostik bezogen, die Funktion wichtiger Randbedingungen, die bei der Interpretation diagnostischer Ergebnisse zu berücksichtigen sind, wie z.B. häufiger Unterrichtsausfall, mehrfacher

Lehrerwechsel oder mangelhafte Ausstattung mit Lehrmitteln (“schulisches Schicksal”; Schmitz, 1964).

10.4 Die Messung pädagogisch-psychologischer Konstrukte

Pädagogisch-psychologische Konstrukte sind kontinuierliche Variablen (Xj), auf denen Schüler in Abhängigkeit von Lehrbemühungen und von ihrer eigenen Aktivität voranschreiten. Das Voranschreiten der Schüler ist auf Indikatorvariablen (Xj’) abzulesen, durch die die (Xj) operational definiert sind. Im Falle von Tests, grundsätzlich aber auch bei Klassenarbeiten, steht (Xj’) für ein Lehrziel oder eine Hierarchie inhaltlich zusammengehöriger Lehrziele, z.B. “Dreisatzaufgaben lösen können”, “Landkarten lesen können” oder “Interpunktionsregeln beherrschen”. Es ist zweckmäßig, die Lernschritte oder die Positionen, die auf der Variablen meßgenau unterschieden werden sollen, durch eine hinreichende Anzahl entsprechender, untereinander gleichwertiger Aufgaben zu repräsentieren.

Die pädagogischen Merkmale, die auf diese Weise erfaßt werden sollen, unterscheiden sich natürlich im Umfang des Kontinuums, d.h. im Hinblick auf die didaktisch unterscheidungsbedürftigen Meßpunkte. So ist z.B. die “Fähigkeit zur Zehnerüberschreitung bei Addition und Subtraktion” im Mathematik-Anfangsunterricht ein relativ “schmales” Merkmal; die “Lesefertigkeit” oder die “Beherrschung der wichtigsten Rechtschreibregeln” sind Beispiele für “breitere” Merkmale. Außerdem können so definierte Merkmale an jeder Stelle dichotomisiert werden, so daß sich jeweils für einen bestimmten Zeitpunkt die Anteile der Schüler ermitteln lassen, die sich ober- bzw. unterhalb des Schnittpunkts befinden (“kriteriumsorientierte” Leistungsmessung; vgl. Abschnitt 6.2). Solche Zwischenkontrollen des Unterrichtserfolgs können bei umfangreicheren Lehreinheiten von didaktischem Interesse sein.

Wie genau Lernzustände der Schüler diagnostiziert werden können, hängt von der Präzision der didaktischen Planung des Unterrichts und von dessen tatsächlichem Verlauf ab. Die Planung richtet sich (anteilig) darauf, aus der Menge der möglichen (Fein-) Lehrziele jeweils eine Auswahl zu treffen, d.h. die Lehrziele so festzulegen und zu ordnen, daß sie der Sachstruktur des Gegenstands, den Ist-Zustanden der Lernenden und deren aktivierbarer Lernkapazität möglichst gut entsprechen.

Im Hinblick auf den üblichen Schulunterricht sind Lehrziele genau dann richtig definiert und damit “curricular zulässig”, wenn empirisch gezeigt werden kann, daß mit angemessenem pädagogischen Aufwand innerhalb angemessener Zeit der Zustand 0 (keiner der Adressaten hat das Lehrziel erreicht) in den Zustand 1 (möglichst alle haben es erreicht) überführbar ist. Was hier als angemessen zu gelten hat, wird anhand von Erfahrungswerten und didaktischem Regelwissen durch pädagogische Experten festgelegt und in die Richtlinien und Lehrpläne für die Schulstufen und Unterrichtsgebiete aufgenommen, z.B. “Strukturierung des Leselehrgangs im 1. Schuljahr”, “jahrgangsspezifischer Aufbau des Rechtschreibunterrichts in der Grundschule” oder “lernen, sich in der näheren und weiteren Umgebung zurechtzufinden” bei Geistigbehinderten.

Für jedes didaktisch zulässige Lehrziel existieren also auf seiten der Lernenden ein Ausgangs-, ein Übergangs- und ein End- oder Soll-Zustand. Den Übergangszustand möglichst kurz zu halten, ist ein wichtiger und althergebrachter pädagogischer Grund-

satz (Ökonomieprinzip; Comenius [vgl. 1959, §3]: “Richtig lehren bedeutet bewirken, daß jemand schnell, angenehm und gründlich lerne”). Unter jeweils vergleichbaren Bedingungen gilt dabei näherungsweise: Je breiter der Ausschnitt aus einem Lehrprogramm (je komplexer das Lehrziel), desto mehr Zeit ist nötig, bzw. je schmaler der Ausschnitt (je elementarer das Lehrziel), desto kürzer ist die Übergangszeit.

10.5 Die diagnostischen Parameter

Dem vorigen Abschnitt zufolge geht die empirische Verteilung der Meßwerte von (Xj’) auf folgende Bedingungen zurück:

(a) den vorgegebenen Lehrzielbereich, d.h. auf die Breite, die Lage und das Abstraktionsniveau des Ausschnitts aus dem Merkmalskontinuum von (Xj), der durch das Meßverfahren repräsentiert wird
(b) den Meßzeitpunkt
(c) die unterschiedliche Lernfähigkeit und die aktuelle Befindlichkeit der zugelassenen Schüler
(d) die unterschiedliche curriculare Validität des erteilten Unterrichts
(e) die Meßungenauigkeit des Instruments (den Meßfehler, der auf mangelnder innerer Konsistenz beruht).

Für jede an Lehrzielen orientierte Diagnostik besteht also ein Problem darin, die Breite und die Lage des Ausschnitts aus der Lehrzielmatrix festzulegen, der für ein bestimmtes Schülerkollektiv durch ein bestimmtes Verfahren abgedeckt werden soll. Ein weiteres diagnostisches Problem liegt in der Lokalisierung der Meßpunkte auf dem Zeitkontinuum, d.h. in bezug auf die korrespondierenden Teilausschnitte aus der Unterrichtssequenz. Für Meßzeitpunkte, zu denen sich mindestens ein Schüler im Hinblick auf mindestens eins der aufgenommenen Lehrziele schon oder höchstens noch im Übergangszustand befindet, ist Streuung zwischen den Individuen zu erwarten; die Größe der Streuung hängt unter sonst gleichen Bedingungen vom Meßzeitpunkt ab.

Da die statistische Aufgabenschwierigkeit durch den Meßzeitpunkt mitbestimmt wird, sind die Meßwertverteilungen umso empfindlicher gegen die Wahl der Meßzeitpunkte, je schmaler der Ausschnitt aus der Lehrzielmatrix ist, bzw. je kürzer die Übergangszeit. Entsprechend sind Verfahren mit einem breiteren Lehrzielspektrum robuster gegen die Lage der Meßzeitpunkte, d.h. sie liefern länger didaktisch verwertbare Information. Die Abbildung 10.2a veranschaulicht die Verteilungen einer Schulleistungsvariablen (X), die zu verschiedenen Zeitpunkten (t) mit einem Breitbandverfahren gemessen wird. Für jeweils einen Zeitpunkt gibt die Verteilung zugleich die relative Schwierigkeit der Aufgaben wieder. Die Zunahme der Lösungswahrscheinlichkeit für homogene Aufgaben, die dasselbe (Teil-)Lehrziel, d.h. denselben Meßpunkt (X_{crit}) repräsentieren, verdeutlicht Abbildung 10.2b (nach Tent & Waldow, 1984).

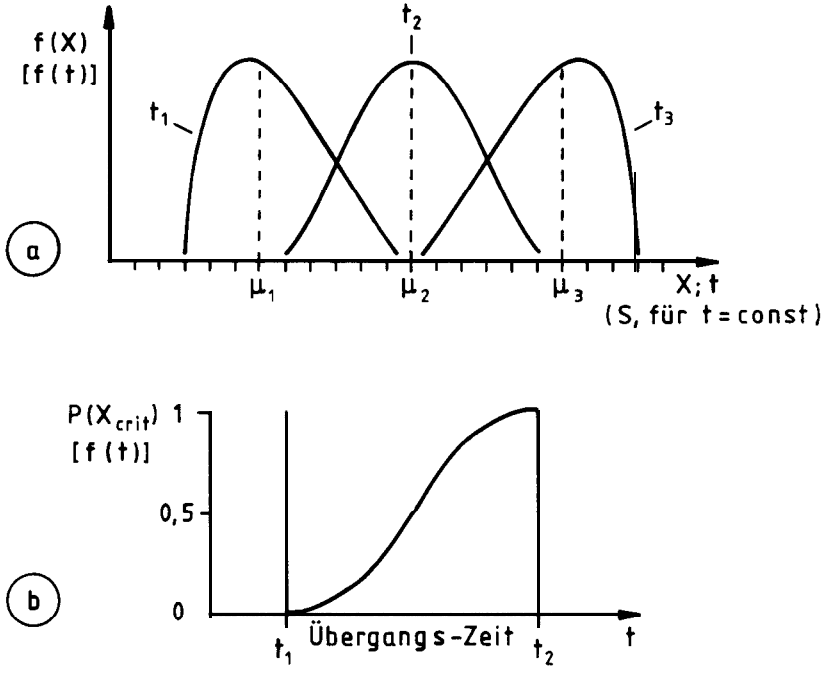


Abbildung 10.2: Verteilung einer Schulleistungsvariablen (X) zu verschiedenen Zeitpunkten (t) (S = Schwierigkeit; 10.2.a) und Zunahme der Lösungswahrscheinlichkeit für einen Kriteriumswert bei lehrzielhomogenen Aufgaben (10.2.b).

Wegen unterschiedlicher Anregungsbedingungen, denen die Schüler außerhalb der Schule ausgesetzt sind, wegen Überschneidungen innerhalb der Unterrichtssequenzen und wegen unterschiedlicher Lern- und Vergessenseffekte erscheint es realistisch anzunehmen, daß für ein bestimmtes Schülerkollektiv für nahezu jeden Meßzeitpunkt (Teil-)Lehrziele angegeben werden können, für die jeweils gilt, daß sich alle Schüler im Zustand 0, alle im Zustand 1, (fast) alle im Übergangszustand (Ü), einige noch im Zustand 0 (die übrigen in Ü oder l), einige schon im Zustand 1 (die übrigen in Ü oder 0) und einige im Übergangszustand (die übrigen in 0 oder 1) befinden. Der didaktische Nutzen eines diagnostischen Verfahrens wird im allgemeinen umso größer sein, je genauer es die verschiedenen Lernzustände der Schüler abzubilden gestattet. Dies angemessen zu beachten, ist das zentrale Problem der Konstruktion von Schulleistungtests. Angemessen heißt hier, Verfahren zu entwickeln, die Meßökonomie und didaktische Ergiebigkeit optimal in sich vereinen. Verschiedene Modelle einer "Lehrplanorientierten Diagnostik" werden bei Shapiro & Terr (1990) vorgestellt.

Die neueren Modellansätze zur Präzisierung des diagnostischen Vorgehens im Sinne "maßgeschneiderter" individualisierter Mikrostrategien (vgl. Abschnitt 8) erfordern einen sehr hohen Konstruktionsaufwand. Sie setzen für jeden Lehrzielausschnitt und jede Schülerpopulation umfassende und vollständig durchprogrammierte Sätze reliabler und valider Testaufgaben voraus. Für die diagnostische Praxis in der Schule spielen sie vorläufig noch keine wesentliche Rolle.

Beschränkt man sich bei der Festlegung der Breite und Lage des Ausschnitts aus einer Lehrzielmatrix auf komplexe Lehrziele oder Lehrziele höheren Abstraktionsniveaus, erhält man, didaktisch gesehen, eher grobmaschige Verfahren. Prognostisch verwendet, enthalten sie, wie manche der konventionellen Schulleistungstests, für die auf Mikrolehrziele gerichteten didaktischen Einzelakte nur wenig Information. Sie liefern lediglich, nach Maßgabe des zu erwartenden Lerntransfers, allgemeine Trends der Erfolgswahrscheinlichkeit für verschieden leistungsfähige Schülergruppen.

Wollte man sich auf Mikrolehrziele, im Grenzfall auf kleinste didaktische Handlungseinheiten und deren Verhaltenskonsequenzen beim Lernenden beschränken, hätte dies eine Kontrolldichte zur Folge, die einen didaktisch flüssigen und reagiblen Unterricht kaum noch zuließe. Komplexität bzw. Abstraktionsniveau der Lehrziele und Kontrolldichte korrelieren: Je umfassender (abstrakter) die Lehrziele, desto größer können die Abstände zwischen den Meßzeitpunkten sein, und entsprechend umgekehrt ("summative" versus "formative" Evaluation im Sinne von Bloom, Hastings & Madaus, 1971). Im einen Fall vermindert sich der unmittelbare didaktische Nutzen nicht nur, weil die Information zu allgemein wird, sondern auch deshalb, weil sie zunehmend zu spät käme. Im anderen Fall würde sich der didaktische Nutzen vermindem, weil diagnostischer Perfektionismus die Gefahr mit sich brächte, seinen lebendigen Gegenstand zu ersticken. Wie jede andere ist auch die Pädagogisch-psychologische Diagnostik nicht Selbstzweck. So grundlegend ihre Bedeutung auch ist, sie untersteht dem bereits früher erwähnten Primat der Didaktik. Wir stoßen hier auf das Bandbreiten-Genauigkeits-Dilemma (Cronbach, 1970, S. 179-182), das als Variante des Generalisten-Spezialisten-Dilemmas verstanden werden kann: Man erfahrt entweder immer weniger Genaues über immer mehr (im Grenzfall "nichts über alles") oder immer Genaueres über immer weniger (im Grenzfall "alles über nichts"; Tent & Waldow, 1984, S. 17-19).

10.6 Meßdichte und didaktische Ergiebigkeit

Auch unabhängig von der Anwendung formeller Testverfahren kennen erfahrene Lehrer das Problem der Verhältnismäßigkeit des pädagogischen Kontrollaufwands, d.h. der Relation der Häufigkeit und Gründlichkeit von Lernkontrollen zu deren didaktischen Nutzen (Problem der Kontrolldichten-Optimierung; vgl. Kaminski, 1982, zur Taxonomie psychodiagnostischer Prozesse).

Nach den vorangegangenen Darlegungen sind aus didaktischer wie aus diagnostischer Sicht tendenziell mittlere Testzeitabstände geboten, d.h. mittlere Dichten der formalisierten Rückkoppelung (s. Abbildung 10.3). Bei mittlerer Meßdichte ist unter sonst gleichen Bedingungen hypothetisch die höchste didaktische Ergiebigkeit zu erwarten. Das bedeutet, daß es einen Bereich optimaler Kontrolldichte gibt, dessen Überschreitung keinen zusätzlichen Informationsgewinn mehr bringt, sondern u.U. zu einem Abfall führt, weil es mit zunehmender Dichte der Messungen zu unerwünschten Nebenwirkungen und Störeffekten kommen kann. Was jeweils "mittlerer Testzeitabstand" oder "optimale Kontrolldichte" ist, wird von Gegenstand zu Gegenstand sowie mit dem Alter und dem kognitiven Entwicklungsstand der Schüler variieren (vgl. dazu Ingenkamp, 1975, S. 84-87). Didaktische Ergiebigkeit ist das Ausmaß, um das sich die Erfolgswahrscheinlichkeit didaktischer Entscheidungen durch diagnostische Information vergrößern läßt.

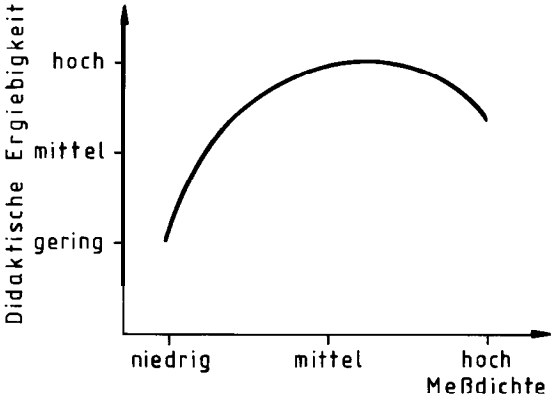


Abbildung 10.3: Hypothetische Beziehung zwischen Meßdichte und didaktischer Ergiebigkeit bei der Anwendung von Schulleistungstests unter sonst gleichen Bedingungen (nach Tent & Waldow, 1984).

Dies entspricht dem Netto-Nutzen einer zureichenden Diagnostik. Eine Diagnostik ist in dem Maße zureichend, wie siefür den Zweck, zu dem sie verwendet werden soll, nachweislich effizient ist, d.h. wie sich mit ihrer Hilfe ökonomischer und/oder wirksamer als mit alternativen Strategien Risiken mindern oder Erfolgswahrscheinlichkeiten erhöhen lassen.

Was jeweils zureichend ist, hängt von der Fragestellung und der Zielsetzung ab. Kommt es z.B. darauf an, Unterschiede im aktuellen Leistungsstand innerhalb einer Schulklasse zu erfassen, genügt vielfach das Urteil erfahrener Lehrer (vgl. z.B. Tent, Fingerhut & Langfeldt, 1976). Es wäre unnötig und unökonomisch, dafür routinemäßig Verfahren mit der höchstmöglichen Meßgenauigkeit einzusetzen. Geht es aber um längerfristige Behandlungszuweisungen, liefert das Lehrerurteil allein erwiesenermaßen keine zureichende Diagnostik. Es müssen, vor allem wenn es sich um Entscheidungen von erheblicher Tragweite handelt, formalisierte Meßverfahren hinzukommen, die hohen Ansprüchen an die üblichen Testgütekriterien genügen (nach Tent & Waldow, 1984).

Für einen gegebenen Testzeitpunkt (t_i) gilt demnach: Es ist ein mittelgroßer didaktischer Ausschnitt, und damit ein mittellanger zeitlicher Abschnitt, diagnostisch valide abzudecken. Sollen dabei interindividuelle Unterschiede vollständig erhoben werden, muß sich, bei zeitlich richtiger Plazierung, unter den verschiedenwertigen Lehrzielen einer vorgegebenen Lehrzielmatrix mindestens eines befinden, für das der Grenzfall 0 und mindestens eines, für das der Grenzfall 1 gilt; die anderen hätten irgendwo dazwischen zu liegen. Das gleiche gilt, wenn es darum geht, die curriculare Validität von Unterricht unter sonst gleichen Bedingungen genau in Erfahrung zu bringen.

Wieweit es sich bei den pädagogisch-psychologischen Konstrukten, die hier unter dem Sammelbegriff “Schulleistung” subsumiert werden, um Konstrukte im Sinne psychologischer Fähigkeiten oder Eigenschaften, d.h. um Persönlichkeitsmerkmale i.e.S. handelt, wäre im Bedarfsfall gesondert zu untersuchen. Es geht dabei um die inhaltliche Bedeutung der retestreliablen Varianzanteile an den Ergebnissen der wie

auch immer etikettierten Prüfverfahren. Ein Schulleistungstest, für den dies nachgewiesen ist, liefert damit - über den Ist-Zustand hinaus - zusätzliche diagnostische Informationen, die pädagogisch umso nützlicher sein können, je valider er relevante Merkmale abbildet. So kann z.B. ein Mathematiktest neben der manifesten Schulleistung in einem definierten curricularen Ausschnitt anteilig Rechnerisches oder Schlußfolgerndes Denken als Subdimensionen der Intelligenz miterfassen. Unter sonst gleichen Bedingungen kann die Meßdichte umso niedriger angesetzt werden, je retestreliabler, d.h. je zeitstabiler oder behandlungsresistenter ein betrachtetes, didaktisch valides Merkmal ist. Ungeachtet ihrer Schwächen eignet sich die klassische Testtheorie besonders gut für die Analyse, welche der zahlreichen Aspekte von Schulleistung zeitlich weniger stabil sind, bzw. welche als "stabiler" gelten können und mit welchen personalen und/oder situativen Randbedingungen dies korreliert (Tent, Fingerhut & Langfeldt, 1976; Tent, 1991).

Ein Merkmal, für das überhaupt keine retestreliablen Meßwerte zu erheben sind, ist didaktisch wertlos. Didaktisch besonders interessant sind solche Variablen, für die empirisch gezeigt werden kann, daß sich infolge pädagogischer Intervention nicht nur die Mittelwerte und Streuungen sondern auch die Retestkoeffizienten ändern für die also systematische Wechselwirkungen zwischen Behandlung und Schülermerkmal bestehen. Beispiele dafür wurden bereits in Abschnitt 10.3 erwähnt (Aptitude-Treatment-Interaction, ATI).

10.7 Nebenwirkungen und Fehlerquellen

In der diagnostischen Praxis ist mit einer Reihe von Problemen zu rechnen, die man beachten muß, wenn man sachgerecht vorgehen und die Befunde angemessen interpretieren will. Es handelt sich dabei um nachteilige Nebenwirkungen, die der Einsatz diagnostischer Verfahren mit sich bringen kann, sowie um Mängel und Fehler, die bei der Aufnahme, der Speicherung und der Verwertung diagnostischer Informationen auftreten können (s. Kasten S. 221).

10.7.1 Problematische Nebenwirkungen

(a) Rückbindungseffekte

Wegen der notwendigen Verschränkung von Unterricht und Schulleistungsdiagnostik kann es insofern zu Rückwirkungen auf den Unterrichtsablauf und das Lernverhalten der Schüler kommen, als übermäßig stark auf die Prüfungen hin gelehrt und gelernt wird. Das mag zwar im Sinne der Standardisierung der Bedingungen nicht unerwünscht sein, führt aber dann zu pädagogisch fragwürdigen Zwangen, wenn deshalb auf die Flexibilität verzichtet werden müßte, im Unterricht aktuelle Anlässe aufzugreifen oder sich auf spontane Bedürfnisse bei Schülern und Lehrern einzustellen. Bis jetzt ist allerdings nur wenig darüber bekannt, in welchem Maße in Schulen tatsächlich auf Prüfungen hin unterrichtet wird, und wie sich dies im einzelnen auswirkt (Ingenkamp, 1975; Kellaghan et al., 1982). Auch aus dieser Sicht wird deutlich, daß es - wie schon bei der Wahl der Meßzeitpunkte - nur um eine mittlere Kontrolldichte gehen kann. Eine inhalt-

Übersicht über mögliche Nebenwirkungen und wichtige Fehlerquellen

Die Anwendung diagnostischer Verfahren in der Schulpraxis kann zum einen pädagogisch bedenkliche Folgen nach sich ziehen. Zum anderen kann die Güte der Urteilsbildung - in erster Linie bei den "weichen" Verfahren - durch systematische Fehler beeinträchtigt sein. In der Literatur werden zahlreiche Fehlerquellen beschrieben und nach unterschiedlichen Gesichtspunkten geordnet. Die hier behandelte Auswahl faßt für die Schulpraxis wichtige Aspekte zusammen:

(1) ***Problematische Nebenwirkungen***
- (a) Orientierung des Unterrichts an Prüfungen (Rückbindungseffekte)
- (b) Leistungsdruck und Belastung des sozialen Klimas (Sozialpsychologische Effekte)
- (c) Fragwürdige Erfolgs- und Mißerfolgszuschreibung (Allgemeine Attribuierungsprobleme)

(2) Inferenzfehler und Einstellungseffekte
- (a) Fragwürdige Ursachenerklärung (spezielle Attribuierungsfehler)
- (b) Selektive Informationsaufnahme (Erwartungseffekte)

(3) Theoriefehler
- (a) Fragwürdige Deutung dominanter Merkmale (Haloeffekte)
- (b) Fragwürdige Persönlichkeitstheorie (Logische Fehler)

(4) Erinnerungs- und Urteilsfehler
- (a) Reihenfolgewirkungen bei der Informationsaufnahme (Positionseffekte)
- (b) Akzentuierung von Unterschieden (Kontrasteffekte)
- (c) Einengung des Urteilsspektrums (Urteilstendenzen, Referenzfehler)

Wie andere, ist auch dieser Ordnungsansatz nicht überschneidungsfrei.

lich vollständige diagnostische Abdeckung anzustreben, wäre didaktisch von vornherein widersinnig.

Die Gefahren der Rückbindung des Unterrichts an die Diagnostik lassen sich in der Regel durch eine bedachtsame Dosierung der Kontrollen in pädagogisch vertretbaren Grenzen halten. Soweit die Zahl und die zeitliche Staffelung etwa von Klassenarbeiten nicht vorgeschrieben sind, bleibt das, was in diesem Zusammenhang angemessen ist, Faustregeln und dem pädagogischen "Augenmaß" der Lehrer und Schulpsychologen überlassen. Für die deutschen Verhältnisse direkt verwertbare Forschung gibt es dazu kaum (vgl. Schneider, 1987).

(b) Sozialpsychologische Effekte

Der pädagogisch-psychologischen Leistungsmessung wird gelegentlich vorgehalten, sie erzeuge Konkurrenzverhalten und Gruppendruck zwischen den Schülern, sie hemme den Lernfortschritt durch Stigmatisierung und Auslösung von Angstgefühlen, sie fördere unlautere Praktiken (Mogeln) und trage zur Neurotisierung einzelner Schüler sowie des pädagogischen Klimas bei. Solche Begleiterscheinungen, die besonders der gruppennorm-orientierten Leistungsrückmeldung angelastet werden, sind natürlich ernst zu nehmen, wo sie auftreten. Welche Bedeutung ihnen in der Schulpraxis insgesamt zukommt, ist schwer einzuschätzen, weil es auch dazu an verwertbaren Forschungsergebnissen mangelt.

Wie Schulnoten mit der Rückmeldung über Leistung und Verhalten auch als Anreiz wirken sollen, können und sollen Testergebnisse selbstverständlich den motivationalen Zustand der Adressaten beeinflussen: bei den Schülern, den Lehrern und den Eltern. Die nachteiligen Nebeneffekte, die von Noten wie von Testergebnissen ausgelöst werden können, gehen nicht auf die diagnostische Information oder Methodik als solche zurück; sie ergeben sich erst aus dem Verwertungszusammenhang (wie z.B. beim Numerus clausus), bzw. aus dem pädagogischen Umgang mit der Information in Schule und Elternhaus. Sie sind kein stichhaltiges Argument gegen eine valide Diagnostik. Unerwünschte Begleiterscheinungen dieser Art zu vermeiden oder aufzufangen, ist daher nicht primär ein diagnostisches sondern ein (i.e.S.) pädagogisches, u.U. ein bildungspolitisches Problem.

So sollten Lehrer beispielsweise bei schwächeren Schülern einseitig gruppennorm-orientierte Leistungsruckmeldung vermeiden (Rheinberg, 1980). Angesichts der vielen Bedingungen, die die Wirksamkeit von Lob und Tadel beeinflussen, sind andrerseits aber auch "paradoxe" Wirkungen zu beachten, die vor allem gutgemeintes Lob nach sich ziehen kann, wenn nämlich bei Schülern der Eindruck entsteht, sie würden gelobt, weil sie es ihrer Begabungsschwäche wegen besonders nötig hätten (W.U. Meyer, 1984; Meyer et al., 1988; Rheinberg & Weich, 1988; Blickle, 1991). Diese Art der Sanktionsverarbeitung, die man nach der Berliner Redensart als Nachtigall-Effekt bezeichnen kann, kommt vermutlich seltener vor, als zunächst angenommen (Hofer & Pikowsky, 1988). Sie wird sich am ehesten durch eine wohlwollend-sachliche und verständnisvolle, aber von persönlicher Sympathie-/Antipathiebekundung freie Rückmeldung im Sinne von Tausch & Tausch (1973) umgehen lassen.

(c)Allgemeine Attribuierungsprobleme

Lehrzielorientierte diagnostische Verfahren sind banalerweise curricular umso valider, je enger die Orientierung an den Lehrzielen ist, deren Realisierung sie kontrollieren sollen. Die üblichen von Lehrern konzipierten Klassenarbeiten haben den Vorteil, daß sie dem tatsächlichen Unterrichtsverlauf angepaßt werden können. Dies ist bei standardisierten Testverfahren in der Regel nicht möglich. Dem Vorzug der größeren Flexibilität der Klassenarbeiten steht der Nachteil gegenüber, daß ihre instrumentelle Güte nicht bekannt ist. Nachteil der meisten verfügbaren Tests ist ihre Starrheit. Sie können weder die didaktischen Freiräume des Lehrers bei der Unterrichtsgestaltung vollständig berücksichtigen, noch unvorhergesehene Abweichungen vom geplanten Verlauf des Unterrichts auffangen. Sie erfassen Schulleistungen auf dem Niveau von Grobzielen in Form von Fakten- und Regelwissen und auf Feinzielebene durch eine repräsentative Auswahl von Aufgaben, die zu ihrer Lösung nicht nur Behalten voraussetzen, sondern einen Lerntransfer erfordern, weil sie in genau dieser Form im Unterricht nicht vorgekommen sind.

Schulleistungstests lassen jedoch valide Aussagen über Schüler höchstens in dem Maße zu, wie der erteilte Unterricht seinerseits als lehrzielvalide gelten kann. Dies ist zu beachten, wenn die Testergebnisse, wie üblich, den Schülern attribuiert werden sollen. Lehrziele und Lehreffekte müssen auseinandergehalten werden, weil sie auf verschiedene Weise divergieren können. Im Verhältnis der erklärten Lehrziele zum realisierten Unterricht sind theoretisch fünf Fälle möglich:

(a) Das Lehrziel ist im realisierten Unterricht überhaupt nicht enthalten
(b) das Lehrziel ist teilweise realisiert
(c) das Lehrziel ist teilweise, außerdem ist noch anderes realisiert

(d) das erklärte Lehrziel und der realisierte Unterricht stimmen vollständig überein und
(e) das Lehrziel ist vollständig, daneben ist noch anderes realisiert.

Bei den "Fremdanteilen" kann es sich um implizite informelle Aspekte des Unterrichts ("heimlicher Lehrplan") oder um nachträgliche Zusätze handeln. Bereits bei der empirischen Validierung und Normierung von Schulleistungstests muß die Lehrzielvalidität des Unterrichts gewährleistet sein, an dem die Analyse vorgenommen wird (vgl. Tent & Waldow, 1984). Werden Testergebnisse zu wichtigen Entscheidungen über Schüler herangezogen, ist jeweils "nach bestem Wissen" abzuwägen, wie gut der Unterricht didaktisch gelungen ist, und wieweit er etwa durch äußere Umstände, z.B. Unterrichtsausfall oder Störungen, belastet war.

10.7.2 Inferenzfehler und Einstellungseffekte

(a) Spezielle Attribuierungsfehler

Ereignisse, Verhalten und Leistungen auf Ursachen zurückzuführen, entspricht offenbar einer tief verwurzelten Denkgewohnheit. Es erscheint uns selbstverständlich, Schulleistungen kausal zu betrachten. Vereinfacht dargestellt, spielen bei der subjektiven Erklärung des Zustandekommens der Leistung von Personen nach Weiner (vgl. 1984) vor allem vier Klassen von Determinanten eine Rolle. Sie unterscheiden sich zum einen danach, ob sie der Person oder den äußeren Umständen zugeordnet werden, zum anderen danach, ob sie als eher stabil oder eher variabel gelten (Tabelle 10.1).

Tabelle 10.1: Schema der subjektiven Determinanten personaler Leistungen (in Anlehnung an Weiner, 1984, S. 270).

	Lokalisierung	
Stabilitätsgrad	intemal	extemal
eher stabil	Eigenschaften Fähigkeiten "Begabung"	(vermutete oder tatsächliche) Aufgabenschwierigkeit
eher variabel	Aktuelle Motivation (Anstrengung, Stimmungslage) Gesundheitszustand	"Zufall" (Glück, Pech)

Dieses Attribuierungsschema kann nicht nur - sofern sie alt genug dazu sind - von Schülern genutzt werden, um sich und anderen Erfolg oder Versagen in der Schule zu erklären, es kommt auch dem Erklärungsmuster entgegen, mit dem Lehrer auf verschiedenen Abstraktionsebenen versuchen, sich Schülerleistungen verständlich zu machen (vgl. Hofer, 1986, Kap. 7 und 8). Es liegt auf der Hand, daß dabei immer dann mit Fehlschlüssen gerechnet werden muß, wenn die Qualität der Beobachtungsdaten, auf die man sich stützt, zu wünschen übrig läßt oder unbekannt ist. Dies betrifft ins-

besondere die Beurteilung mündlicher Leistungen und des Sozialverhaltens. Die Folgen können umso bedenklicher sein, je höher das Abstraktionsniveau der Inferenz. Die Gefahr besteht vor allem darin, Schülern ohne ausreichende Grundlage abstrakte Fähigkeiten und Eigenschaften, also Persönlichkeitsmerkmale i.e.S., zuzuschreiben und damit bestimmte Erwartungen an ihr künftiges Verhalten zu verknüpfen.

Demgegenüber ist daran zu erinnern, daß Schulleistungsdiagnosen zunächst nur deskriptive Informationen über merkmalsspezifische Ist-Zustände liefern. Annahmen über die Stabilität des betreffenden Merkmals und künftige Leistungen sind nur in dem Maße zulässig, wie dies durch wiederholte, situationsübergreifende Beobachtung, bzw., im Fall standardisierter Testverfahren, durch die empirische Retest-Reliabilität und Validität abgedeckt ist. Weitergehende "kausale" Erklärungen, z.B. für Schulversagen, sind allenfalls möglich, soweit sich dafür aus der Vorgeschichte und aus den Lebensumständen des Schülers Anhaltspunkte für begründete Vermutungen ergeben, z.B. Krankheiten, Unfälle, Deprivation oder besondere "Schlüsselerlebnisse" (life events).

(b) Erwartungseffekte

Die Verhaltens- und Leistungserwartungen, die Lehrer aufgrund fehlerhafter Zuschreibungen entwickeln, können sich dadurch verfestigen, daß in der Folge bevorzugt solches Schülerverhalten "wahrgenommen" wird, das der Erwartung entspricht, während andere Verhaltenselemente ausgeblendet werden. Über die Rückmeldung seiner Erwartungen an den Schüler kann der Lehrer u.U. bewirken, daß das erwartete Verhalten tatsächlich vermehrt gezeigt wird. Man spricht hier von "sich selbst erfüllenden Vorhersagen" (self-fulfilling prophecies), die auf der Grundlage selektiver "Person-Wahrnehmung" oder, genauer, auf der Grundlage einseitiger kognitiver Inferenz aus Wahrnehmungsdaten oder Informationen durch Dritte zustandekommen. Dieses Phänomen wird im Anschluß an Rosenthal und Jacobson (1968, 197 1; Rosenthal, 1975) auch Pygmalion-Effekt genannt. Es kann sich nicht nur auf Leistungsvariablen sondern auch auf Verhaltenstereotype und "charakterliche" Merkmale erstrekken (vgl. Ludwig, 1991). An diagnostische Artefakte dieser Art ist besonders zu denken, wenn die Konsequenzen für den Betroffenen gravierend sind, etwa bei präventiven Risikoprognosen (Krapp, 1986, S. 628-630). Nach der Definition von Lernbehinderung ist es z.B. zulässig, Kinder bereits in die Sonderschule einzuweisen, wenn ein längerdauemdes und umfassendes Versagen in der Regelschule zu erwarten ist. Dabei ist in diesem Fall fraglich, ob die Sonderschule überhaupt eine pädagogisch bessere Behandlungs-Alternative darstellt (Tent, Witt, Zschoche-Lieberum & Bürger, 1991).

Der Pygmalion-Effekt kann als Spezialfall der allgemeinen Tendenz zur Bestätigung von Hypothesen bei der Urteilsbildung gelten (confirmation bias). Sie begünstigt die hypothesenkonsistente Auswahl und/oder Verarbeitung von Informationen und erhöht dadurch die Wahrscheinlichkeit, eine Hypothese als bestätigt anzusehen (Hager & Weißmann, 1991).

In welchem Maße tatsächlich mit Pygmalion-Effekten zu rechnen ist, läßt sich kaum vorhersagen, weil die Entstehungsbedingungen vielfaltig sind und wenig einheitlich erscheinen. Sie treten vermutlich umso eher auf, je intensiver die Erwartung ist und je weniger andere Informationen neben den erwartungsweckenden zur Verfü-

gung stehen (zur Forschungslage vgl. Hofer, 1986, Kap. 8; Chow, 1990). Obwohl man nicht davon ausgehen muß, daß diese Effekte stark verbreitet sind, sollten diagnostische Aussagen hohen Abstraktionsniveaus auch deswegen nur gemacht werden, wenn man sich auf dafür geeignete Verfahren stützen kann. Fehlt es daran, beschränkt man sich besser auf die Wiedergabe des beobachteten Verhaltens (z.B. statt "Peter ist ein Lügner": "Peter hat bei dieser und jener Gelegenheit dieser oder jener Person gegenüber aus diesem oder jenem Grund die Unwahrheit gesagt oder verschwiegen, was er wußte"). Die verbale Fassung ist zwar umständlicher als die nominale, entspricht aber der Erkenntnislage.

10.7.3 Theoriefehler

Fehler dieser Art beruhen auf ungeprüften Annahmen über psychologische oder logische Zusammenhänge zwischen Merkmalen. Sie werden daher auch als Korrelationsfehler bezeichnet (Kleiter, 1973). Die Annahmen können Bestandteil naiver (expliziter oder impliziter) Theorien über Aufbau und Funktionieren der Persönlichkeit sein.

(a) Halo- oder Hof-Effekte

Von Haloeffekt wird gesprochen, wenn hinsichtlich einzelner beurteilter Personen ein eindrucksmäßig vorherrschendes ("dominantes") Merkmal die vom Betrachter wahrgenommene Ausprägung anderer Merkmale beeinflußt, so wenn ein Lehrer z.B. bei einem "faulen" Schüler festzustellen glaubt, daß er "desinteressiert", "willensschwach" oder auch "minderbegabt" sei. Unabhängig von der objektiven Sachlage strahlt die Etikettierung "faul" bei der Urteilsbildung auf andere Merkmale aus.

(b) Logische Fehler

Mit dem Halo-Effekt verwandt, aber theoretisch davon zu unterscheiden, ist der sogenannte logische Fehler, der zu vergleichbaren Konsequenzen führt, weil der Beurteiler z.B. aufgrund einer impliziten Persönlichkeitstheorie annimmt, daß bestimmte Merkmale allgemein eng miteinander zusammenhängen (zu den Persönlichkeitstheorien von Lehrern s. Bender, 1985; Hofer, 1986). Unter impliziter Persönlichkeitstheorie wird die Gesamtheit der Annahmen verstanden, die jemand über die Zusammenhänge und die Organisation von Eigenschaften bei anderen Menschen besitzt (Hofer, 1986, S. 71). Nach einer solchen Logik verwandte Merkmale werden dann ohne weitere Beobachtung des Einzelfalls generell ähnlich bewertet. So gibt es z.B. Annahmen über die Koppelung von gutem Aussehen mit Freundlichkeit und Energie oder von Freundlichkeit mit Großzügigkeit und Optimismus. Die volkstümlich-naive Trias "dumm, dreist und gefräßig" ist ein weiteres Beispiel dafür. Es kann aber auch aufgrund struktureller Ähnlichkeit der Unterrichtsfächer angenommen werden, daß z.B. Schüler, die in Mathematik gut sind, auch in Physik überdurchschnittlich abschneiden müßten. - Logische Fehler können z.B. durch die räumliche Nähe vorgegebener Kategorien in Beurteilungsbögen begünstigt werden, d.h. die Korrelation zwischen zwei Kategorien kann höher ausfallen, wenn sie unmittelbar aufeinander folgen, als wenn sie räumlich getrennt sind (Nähe-Effekt, proximity-error).

10.7.4 Erinnerungs- und Urteilsfehler

Soweit diagnostische Urteile auf der Beobachtung von Verhalten und der Einschätzung von Leistungen beruhen, können die Ergebnisse auch durch eine Reihe heterogener Effekte verfälscht sein, die vor allem in der Sozialpsychologie beschrieben worden sind und häufig als Beobachtungs- oder Urteilsfehler zusammengefaßt werden.

Unter den üblichen Bedingungen des Schulunterrichts haben Lehrer fast ständig eine Vielzahl von Schülern gleichzeitig "im Auge zu behalten". Sie müssen große Mengen an Information aufnehmen, speichern und verarbeiten. Natürlich können sie bei weitem nicht alles wahrnehmen und behalten, was für ihre diagnostische Urteilsbildung belangvoll sein könnte. Sie können dem gesamten Verhaltensspektrum ihrer Schüler immer nur Stichproben entnehmen, auch wo sie sich auf den einzelnen konzentrieren. Die Entnahme wird offensichtlich von zahlreichen, nur begrenzt kontrollierbaren Faktoren beeinflußt, wie Vorkenntnisse, Einstellungen, Werte und Erwartungen, Art und Bedeutung der Merkmale, sozialer Kontext und aktuelle Bedürfnisse. Wieweit die Beobachtungs-Stichproben jeweils repräsentativ sind, läßt sich in der Praxis kaum feststellen. Doch hängt von der Qualität der Stichproben die Genauigkeit ab, mit der die didaktischen Entscheidungen des Lehrers der pädagogischen Situation gerecht werden können. Erfahrene Lehrer sind deshalb bemüht, "Stichprobenfehler" der hier gemeinten Art gering zu halten, indem sie sich z.B. zur Regel machen, jeden Schüler in jeder Unterrichtsstunde mindestens einmal zu Wort kommen zu lassen.

Neben der irrepräsentativen Stichprobenentnahme (Beobachtungsfehler i.e.S.) sind hauptsächlich folgende Fehlerquellen zu beachten:

(a) Serielle Positionseffekte

Diese aus der Gedächtnispsychologie bekannten Effekte besagen, daß Informationen, die zu Beginn oder gegen Ende einer Informationsabfolge aufgenommen werden, besser im Gedächtnis haften als die in der Mitte befindlichen (Anfangs- und Endbetonung; primacy-recency-effects). An solche Effekte ist z.B. bei der Bewertung mündlicher Prüfungsleistungen und längerer Beiträge im Unterricht zu denken.

(b) Kontrasteffekte

Hier geht es um eine andere Art von Reihenfolge-Effekt. Vor allem die Urteile über die Leistung in aufeinanderfolgenden mündlichen Prüfungen können davon beeinflußt werden, ob der Prüfung eines Kandidaten eine "gute" oder eine "schwache" Prüfung vorangegangen ist. Der Kontrasteffekt besteht darin, daß die Leistungsunterschiede in den Urteilen starker akzentuiert werden, als objektiv gerechtfertigt wäre (vgl. Birkel, 1978). Grundsätzlich gilt dies auch für die fortlaufende Korrektur schriftlicher Arbeiten.

Im Anschluß an Murray (1938) wird als Kontrastfehler außerdem die Tendenz von Beurteilern bezeichnet, Beurteilten Merkmale oder Merkmalsausprägungen zuzuschreiben, die den eigenen Zügen entgegengesetzt sind.

(c) Urteilstendenzen

Damit ist in erster Linie die Neigung von Beurteilern gemeint, das Urteilsspektrum einzuschränken, d.h. entweder gehäuft günstige oder gehäuft ungünstige Urteile abzugeben, bzw. extreme Urteile zu vermeiden und die mittleren überproportional zu bevorzugen. Man spricht dann von Milde-Effekt (leniency-effect; generosity-error), bzw. von Strenge-Effekt und von der Tendenz zur Mitte (error ***of*** central tendency). Sind in den Urteilen beide Extreme über- und die mittleren Werte unterrepräsentiert, kann eine Tendenz zur Schwarz-Weiß-Malerei (Cronbach, 1970) vermutet werden. Solche systematischen Unterschiede in der Urteilsverteilung können u.a. darauf zurückgehen, daß Beurteiler unterschiedliche Vergleichsmaßstäbe anlegen, d.h. daß sie ihre Urteile an unterschiedlichen Referenzpopulationen orientieren. Man spricht deshalb auch von Referenzfehlern (Kleiter, 1973).

Bei der Verwertung diagnostischer Schätzurteile - dies betrifft fast alle Schulnoten - muß auf derartige Unterschiede in der Beurteilungspraxis etwa zwischen Lehrern, Schulen oder Fächern geachtet werden: So fallen z.B. juristische Staatsexamen (bei hoher Durchfall-Quote) traditionell “schlechter” aus als andere Abschlußprüfungen, in denen ein “Gut” bereits unter dem empirischen Durchschnitt liegen kann. Bekannt sind auch die unterschiedlichen Notenverteilungen bei den Schulfächern Religion und Kunst auf der einen und z.B. Latein und Mathematik auf der anderen Seite.

Wie bei den Attribuierungsfehlern und den Erwartungseffekten sind die Entstehungsbedingungen für die Urteilsfehler vielfältig. Man kann die Fehler zumindest partiell über die Schätzung ihres Beitrags zur Urteilsvarianz korrigieren und diesen Anteil durch Anleitung der Beobachter und Beurteiler zu größerer Wahrnehmungsschärfe und begrifflicher Präzision verringem (Hasemann, 1983; Hager & Weißmann, 1991). Man wird sie aber nicht völlig ausschalten können. Als Bestandteil sozialer Kognition gehen sie unausweichlich in die Bildung der Kategorien ein, in die wir gewohnt sind, Gegenstände, Ereignisse und Menschen einzuordnen. Aufnahme und Verarbeitung von Informationen sind anfällig für Fehler, weil wir die Information zu stark verdichten müssen und wichtige Einzelheiten verloren gehen. Hinzu kommt unsere Neigung, an vorhandenen kognitiven Strukturen festzuhalten. Es sollte aber nicht vergessen werden, daß viele der Urteile, die wir uns über andere bilden, im großen und ganzen zutreffen (Schneider, 1991).

Dies gilt grundsätzlich auch für die kognitiven Prozesse, mit denen Lehrer von beobachtetem Schülerverhalten auf zu Grunde liegende Persönlichkeitsmerkmale schließen. Lehrer gehen im Unterricht durchaus differenziert auf die unterschiedlichen Verhaltensmuster ihrer Schüler ein. Allerdings ist die Genauigkeit, mit der sie ihre Eindrücke bilden, bei Persönlichkeitsaspekten geringer als bei Leistungsaspekten, und erwartungsgemäß unterscheiden sich Lehrer danach, welche Merkmale sie besser und welche sie weniger gut einschätzen können (Hofer, 1986, Kap. 3,5 und 6; Dobrick & Hofer, 1991).

Zusammenfassung

In der pädagogisch-psychologischen Diagnostik geht es im wesentlichen um die Feststellung inter- und intraindividueller Unterschiede auf Merkmalen, die sich dem Konstrukt "Schulleistung" subsumieren lassen. "Schulleistung" ist die Resultante aus dem Zusammenwirken einer Vielzahl von Schüler- und Schulmerkmalen. Unter Leistung wird dabei jedes Ergebnis menschlichen Handelns verstanden. Das Konstrukt ist operational über die Indikatorvariablen "Lehrerurteile" und "Testwerte" definiert, die in der Regel Schülern attribuiert werden. Es umfaßt alle Einzelmerkmale, die inhaltlich den verschiedenen Lehrzielen zugeordnet werden können und sich in Abhängigkeit von Lehrbemühungen verändern. Im Hinblick auf die Veränderungen wird zwischen Ausgangs-, Übergangs- und Endzustand unterschieden. Die Meßergebnisse sind stets Ist-Werte. Die Genauigkeit, mit der Lernfortschritte diagnostiziert werden können, hängt von der Präzision der Planung und dem Verlauf des Unterrichts sowie von der Repräsentativität der Testaufgaben bzw. der Beobachtung ab. Die Verteilung der Meßwerte wird von der curricularen Validität des Unterrichts und vom Meßzeitpunkt mitbestimmt. Für die Wahl aufeinanderfolgender Meßzeitpunkte ist anzunehmen, daß eine mittlere Kontrolldichte didaktisch am nützlichsten ist.

Zu starre Handhabung diagnostischer Verfahren kann zu einer pädagogisch bedenklichen Orientierung des Unterrichts an den Prüfungen und damit zu einem Verlust an didaktischer Flexibilität führen. Ungünstige sozialpsychologische Begleiterscheinungen können durch pädagogisch sachgemäßen Umgang mit diagnostischen Ergebnissen vermieden oder aufgefangen werden.

Die Fehler und Störeffekte, die vor allem die auf Personwahrnehmung gestützten Eindrucksurteile verzerren können, sind zwar wegen des Handlungsdrucks und der notwendigen Informationsverdichtung nicht ganz zu vermeiden, lassen sich aber durch Training reduzieren.

Einführende Literatur zu 10.7:

Preiser, S. (1979). Personwahrnehmung und Beurteilung. Darmstadt: Wissenschaftl. Buchgesellschaft.

Weiterführende Literatur:

Bierhoff, H.W. (1986). Personenwahrnehmung. Vom ersten Eindruck zur sozialen Interaktion. Berlin: Springer.

Fiske, S.T. & Taylor, S.E. (1991). Social Cognition. New York: McGraw-Hill.

11. Berufsethische und rechtliche Aspekte

1. Welche Gütestandards und ethischen Grundsätze sind bei diagnostischen Maßnahmen zu beachten?
2. Wer ist für die sachgerechte Durchführung verantwortlich?
3. Welche Rechtsvorschriften regeln die Anwendung diagnostischer Verfahren in der Schule?
4. Wie können diagnostische Maßnahmen rechtlich überprüft werden?
5. Was ist bei der Durchführung wissenschaftlicher Untersuchungen in Schulen zu beachten?

Vorstrukturierende Lesehilfe

Die diagnostische Tätigkeit ist ein wesentlicher Bestandteil des professionellen Handelns von Lehrern und Psychologen. Sachlogik und Berufsethos verlangen, daß sie hohen Ansprüchen genügt. Was die Psychologen betrifft, setzt die Berufsordnung ***für*** Psychologen des Berufsverbandes Deutscher Psychologen von 1986 dafür fachliche und ethische Maßstäbe. Sie enthält Richtlinien, an denen sich auch die Pädagogisch-psychologische Diagnostik orientieren sollte.

Darüber hinaus unterliegt die diagnostische Praxis der rechtlichen Bewertung durch Gesetzgeber, Verwaltung und Gerichte. Soweit sie von Lehrern und Psychologen im öffentlichen Dienst vorgenommen wird, gehört die diagnostische Beurteilung des Verhaltens und der Leistungen von Schülern zu den Dienstpflichten und ist damit Teil des staatlichen Verwaltungshandelns, das durch arbeits-, dienst- und beamtenrechtliche Vorschriften weitgehend geregelt ist.

Aber auch da, wo noch keine speziellen Rechtsvorschriften bestehen, ist die Pädagogisch-psychologische Diagnostik in die allgemeine Rechtsordnung eingebunden. Das Grundgesetz für die Bundesrepublik Deutschland von 1949 steckt den Rahmen des diagnostisch Zulässigen ab. Einzelne diagnostische Maßnahmen und darauf gestützte Entscheidungen können gerichtlich überprüft und von den zuständigen Fachaufsichts-Behörden kontrolliert werden. Die empirische Forschung an Schulen wirft besondere pädagogische und rechtliche Probleme auf.

11.1 Berufsethische Anforderungen

Praktische Diagnostik ist kein Selbstzweck. Wie alles menschliche Handeln ist sie stets und selbstverständlich in gesellschaftliche Zusammenhänge eingebunden. Sie

wird von ihnen mitbestimmt und wirkt auf sie zurück. Sie dient unterschiedlichen Zwecken, die sich ihrerseits unterschiedlich bewerten lassen. Auch deshalb bedarf sie der rationalen Begründung. Wo sie mit Wertvorstellungen und Rechtsgütern in Konflikt geraten kann, bedarf sie darüber hinaus der Legitimation. Als ein von Werten und Zwecken bestimmtes Handeln hat sie wie jede Berufsausübung eine ethische Dimension. Sie läßt sich nach Qualitätskriterien beurteilen, die angeben, wie gearbeitet werden soll, was zulässig und sittlich richtig ist, bzw. was als unzulässig oder moralisch verwerflich (wenn nicht ungesetzlich) gilt.

In diesem Sinne sind berufsethische Richtlinien untergesetzliche Normen, die den Berufsangehörigen verpflichten, bestimmte Gütestandards einzuhalten und Regeln für den Umgang mit Personen, Sachen und Informationen zu befolgen. Sie knüpfen an die besonderen Aufgaben und Ziele des Berufs an und leiten aus der damit verbundenen Verantwortung zumeist auch ab, in welchem Geiste, d.h. mit welcher Grundeinstellung und moralischen Gesinnung der Beruf ausgeübt werden soll. Ein bekanntes frühes Beispiel für solche Leitsätze ist der dem Hippokrates zugeschriebene "Eid" der Ärzte (5./4. Jahrh. vor unserer Zeitrechnung). Bei uns gibt es für Lehrer und Psychologen bis jetzt keine allgemein verbindlichen Regelwerke dieser Art. Auf Richtzielebene haben die Lehrer und die im Erziehungsbereich tätigen Psychologen, wie andere Fachleute, dem Wohl des einzelnen und der Gesellschaft zu dienen. Sie haben, plakativ gesagt, einen gesellschaftlichen Auftrag zu erfüllen, und sie stehen in der Verantwortung vor der nachwachsenden Generation.

In der Erziehungswissenschaft gehen die Auffassungen über die pädagogische Ethik sowie über das Berufsethos der Lehrer und dessen Stellenwert auseinander (einführend Wigger, 1990). Da der Lebenslauf und der berufliche Werdegang fast aller Bürger durch pädagogische Entscheidungen wesentlich beeinflußt werden, erscheint es einleuchtend, zur Verwirklichung des Erziehungsauftrags der Schule von Lehrern ein hohes professionelles Ethos zu verlangen (Brezinka, 1986). Nach verbreiteter Ansicht darf sich der Lehrer nicht nur als staatlich alimentierter Stundengeber verstehen. Es wird erwartet, daß er sich in besonderer Weise engagiert und in seinem Beruf mehr sieht als einen "Job wie jeder andere." Ganz gleich wie man zu solchen Forderungen steht, es läßt sich nicht bezweifeln, daß die mehr oder weniger selbstverständlichen Gütestandards für pädagogisches Handeln auch für den diagnostischen Ausschnitt gelten, und zwar unabhängig davon, ob dies in den didaktischen Handreichungen, Curricula und Dienstanweisungen jeweils explizit angeführt wird oder nicht. Diese Maxime ergibt sich schon aus dem Anspruch auf Wissenschaftlichkeit des Vorgehens und aus der immanenten Sachlogik: Wie bereits dargelegt, kann pädagogisches Handeln insgesamt nicht besser sein als die Diagnose der pädagogischen Zustände, auf deren Veränderung es sich richtet.

Auch in der Psychologie werden berufsethische Fragen aus z.T. kontroverser Sicht diskutiert. Die psychologische Diagnostik ist davon besonders betroffen (vgl. z.B. Hartmann & Haubl, 1984; Jäger, 1986). Doch stellt sich die Lage für die Psychologen insgesamt einheitlicher und konkreter dar als bei den Pädagogen. Der Berufsverband Deutscher Psychologen hat 1986 eine von breitem Konsens getragene "Berufsordnung für Psychologen" erlassen können, die an die Stelle der "Berufsethischen Verpflichtungen" von 1967 getreten ist. Soweit sie nicht höherrangige Rechtsvorschriften aufnimmt, bindet sie allerdings nur die im Verband freiwillig zusammengeschlossenen Mitglieder. Anderen Psychologen kann sie als Orientierungshilfe dienen, solange und soweit sie nicht z.B. durch ein Psychologengesetz überholt wird. Sie

enthält eine Reihe allgemeiner Bestimmungen, die die diagnostische Tätigkeit mitbetreffen, und einige, die sich direkt darauf beziehen (s. Kasten).

Auszug aus ***der Berufsordnung für Psychologen***
(Berufsverband Deutscher Psychologen. Bonn, 1986).

I. ***Präambel***

1. ***Beruf***

Die Aufgabe des Psychologen ist es, das Wissen über den Menschen zu vermehren und seine Erkenntnisse und Fähigkeiten zum Wohl des Einzelnen und der Gesellschaft einzusetzen. Er achtet Würde und Integrität des Individuums und setzt sich für die Erhaltung und den Schutz fundamentaler menschlicher Rechte ein. Der Beruf des Psychologen ist seiner Natur nach frei.

2. Verantwortung

Der Psychologe ist verpflichtet, seinen Beruf gewissenhaft auszuüben und dem Vertrauen, das ihm in seiner Berufsausübung entgegengebracht wird, zu entsprechen. Er muß sich stets der sozialen Verantwortung bewußt sein, die sich daraus ergibt, daß seine Tätigkeit dazu geeignet ist, auf das Leben anderer in besonderer Weise einzuwirken. Der Psychologe anerkennt das Recht des Individuums, in eigener Verantwortung und nach seinen eigenen Überzeugungen zu leben, und bemüht sich in seiner beruflichen Tätigkeit um Sachlichkeit und Objektivität. Er ist wachsam gegenüber persönlichen, sozialen, institutionellen, wirtschaftlichen und politischen Faktoren und Einflüssen, die zu einem Mißbrauch bzw. einer falschen Anwendung seiner Kenntnisse und Fähigkeiten führen könnten.

3. Kompetenz

Verantwortliches berufliches Handeln erfordert hohe fachliche Kompetenz. Der Psychologe ist verpflichtet, sich durch Fortbildung über den jeweiligen Stand der Wissenschaft in Kenntnis zu setzen. Er hat sich dabei auch über die für seine Berufsausübung geltenden Vorschriften zu unterrichten. Der Psychologe bietet nur Dienstleistungen an, für deren Erbringung er durch Ausbildung und fachliche Erfahrung qualifiziert ist. Er orientiert sich dabei an wissenschaftlichen und fachlichen Standards und bedient sich entsprechend überprüfter und anerkannter Methoden. Er hält sich an den Grundsatz der wissenschaftlichen Redlichkeit und überprüft den Erfolg seiner Arbeit. Psychologische Aufgaben übernimmt er nur, wenn er die damit verbundenen Verpflichtungen einhalten kann. Aufgrund seiner Kompetenz handelt der Psychologe in psychologischen Sachfragen eigenverantwortlich und selbständig.

(...)

III. ***Stellung zu Klienten/Patienten***

(...)

2. Aufklärungspflicht

Der Psychologe hat seinen Klienten/Patienten über alle wesentlichen Maßnahmen und Behandlungsabläufe zu unterrichten. (...)

IV. ***Stellung zu Kollegen und anderen Berufsgruppen***

(...)

2. Verhältnis zu Angehörigen anderer Berufe

(1) Der Psychologe ist in der Zusammenarbeit mit Angehörigen anderer Berufe loyal, tolerant und hilfsbereit. Er kennt keine standespolitischen Grenzen und arbeitet mit anderen Berufen zusammen.

(...)

(4) Angestellte oder beamtete Psychologen haben bei Begründung eines Dienstverhältnisses auf ihre eigenverantwortliche Berufsausübung hinzuweisen, insbesondere auf die ihnen kraft Gesetzes obliegende Schweigepflicht.

(...)

VII. ***Umgang mit Daten***

Schweigepflicht

(1) Der Psychologe ist verpflichtet, über alle ihm in Ausübung seiner Berufstätigkeit anvertrauten und bekannt gewordenen Tatsachen zu schweigen (§ 203 StGB), soweit nicht das Gesetz Ausnahmen vorsieht oder ein bedrohtes Rechtsgut überwiegt.

(2) Die Schweigepflicht des Psychologen besteht auch gegenüber Familienangehörigen des Klienten/Patienten und gegenüber Vorgesetzten.

(...)

VIII. ***Ausstellung von Gutachten und Untersuchungsberichten***

1. ***Sorgfaltspflicht***

Allgemein gilt, daß die Erstellung und Verwendung von Gutachten und Untersuchungsberichten vom Psychologen größtmögliche Sachlichkeit, Sorgfalt und Gewissenhaftigkeit erfordert. Gutachten und Untersuchungsberichte sind frist- und formgerecht **anzu**fertigen.

2. Transparenz

Gutachten und Untersuchungsberichte müssen für den Adressaten inhaltlich nachvollziehbar sein.

3. Einsichtnahme

(1) Sind Auftraggeber und Begutachteter nicht identisch, kann das Gutachten bzw. der Untersuchungsbericht nur mit Einwilligung des Auftraggebers dem Begutachteten zugänglich gemacht werden.

(...)

Gefordert werden u.a. Achtung der Würde und Integrität des Individuums, gewissenhafte Berufsausübung und Verantwortungsbewußtsein, das Bemühen um Sachlichkeit und Objektivität sowie Wachsamkeit gegenüber Einflüssen, die zum ***Mißbrauch psychologischer Kompetenz*** führen können. Der Psychologe ist zur fachlichen Fortbildung verpflichtet; er hat sich an wissenschaftlichen Standards zu orientieren und entsprechend überprüfte und anerkannte Methoden zu verwenden. Er soll sich an den Grundsatz der wissenschaftlichen Redlichkeit halten, den Erfolg seiner Arbeit kontrollieren und mit Angehörigen anderer Berufe loyal zusammenarbeiten.

Für die Erstellung und Verwendung von ***Gutachten*** und ***Untersuchungsberichten*** werden "größtmögliche Sachlichkeit, Sorgfalt und Gewissenhaftigkeit" gefordert; außerdem müssen sie "für den Adressaten inhaltlich nachvollziehbar" sein (VIII. 1 und 2). Die gesetzliche Verpflichtung des Psychologen zur Wahrung von Privatgeheimnissen ("Schweigepflicht", § 203 StGB). die auch gegenüber Familienangehörigen

und Vorgesetzten besteht, ist ausdrücklich in den Regelkanon aufgenommen (VII. 1; X.2.2; s. Kasten auf dieser Seite). Soweit diagnostische Verfahren zu Forschungszwecken verwendet werden, sind die Richtlinien für die Planung und Durchführung empirischer Forschungsvorhaben zu beachten (X. 1 und 2; zur Ethik der psychologischen Forschung s. Schuler, 1980). Verstöße von Mitgliedern gegen die Berufsordnung können durch das Ehrengericht des Verbandes geahndet werden (XI).

1. ***Verpflichtung des Psychologen zur Verschwiegenheit***

Zusammen mit den Angehörigen einiger anderer Berufe, wie Mediziner, Anwälte, Steuerberater, Erziehungsberater, Sozialpädagogen und Versicherungsmitarbeiter, unterliegen Psychologen der ***Schweigepflicht.*** Ein Verstoß dagegen kann zu strafrechtlicher Verfolgung führen. Das in der Bundesrepublik Deutschland gültige ***Strafgesetzbuch*** (StGB) von 1871 (in der Fassung von 1987) sieht vor:

"§ 203. [Verletzung von Privatgeheimnissen] (Auszug)
(1) Wer unbefugt ein fremdes Geheimnis, namentlich ein zum persönlichen Lebensbereich gehörendes Geheimnis oder ein Betriebs- oder Geschäftsgeheimnis offenbart, das ihm als
1. (...)
2. Berufspsychologe mit staatlich anerkannter wissenschaftlicher Abschlußprüfung, (...)
6. (...)
anvertraut worden oder sonst bekanntgeworden ist, wird mit Freiheitsstrafe bis zu einem Jahr oder mit Geldstrafe bestraft.
(2) (...)"

Die Strafandrohung gilt bereits für Studenten; sie gilt auch nach dem Tod der Verpflichteten und der Verletzten:

"(3) Den in Absatz 1 Genannten stehen ihre berufsmäßig tätigen Gehilfen und die Personen gleich, die bei ihnen zur Vorbereitung auf den Beruf tätig sind. Den in Absatz 1 und den in Satz 1 Genannten steht nach dem Tod des zur Wahrung des Geheimnisses Verpflichteten ferner gleich, wer das Geheimnis von dem Verstorbenen oder aus dessen Nachlaß erlangt hat.
(4) Die Absätze 1 bis 3 sind auch anzuwenden, wenn der Täter das fremde Geheimnis nach dem Tode des Betroffenen unbefugt offenbart."

Das Strafmaß fällt höher aus, wenn der zur Verschwiegenheit Verpflichtete sein Wissen "vermarktet" oder zum Schaden anderer nutzt:

"(5) Handelt der Täter gegen Entgelt oder in der Absicht, sich oder einen anderen zu bereichern oder einen anderen zu schädigen, so ist die Strafe Freiheitsstrafe bis zu zwei Jahren oder Geldstrafe."

Die Verletzung von Privatgeheimnissen wird nicht von Amts wegen (Offizialverfahren) verfolgt, sondern nur auf Antrag des Verletzten oder, nach dessen Tod, auf Antrag der Angehörigen oder Erben (Antragsdelikt; § 205 StGB).

II. ***Kein Recht auf Zeugnisverweigerung im Strafverfahren***

Obwohl die Psychologen zur Verschwiegenheit verpflichtet sind, steht ihnen nach der in der Bundesrepublik Deutschland gültigen ***Strafprozeßordnung*** (StPO) von 1877 (in

der Fassung von 1987) ***nicht generell*** das Recht zu, aus beruflichen Gründen das Zeugnis zu verweigern. Ein solches Recht wird in abschließender Aufzählung u.a. Geistlichen, Anwälten, Steuerberatern, Ärzten, Hebammen, Mitarbeitern von Beratungsstellen nach § 218b StGB, Parlamentsmitgliedern und Journalisten eingeräumt; ***als Berufsgruppe*** sind die Psychologen bislang davon ausgenommen (§ 53 StPO [Zeugnisverweigerungsrecht aus beruflichen Gründen]).

III. ***Verpflichtung zur Amtsverschwiegenheit***

Personenbezogene Daten sind im öffentlichen Dienst auch durch das ***Amtsgeheimnis*** (Dienstgeheimnis) geschützt (Beamtengesetze; BAT). Strafrechtlich ist die Amtsverschwiegenheit mit derselben Strafandrohung wie in Absatz 1 durch § 203 Absatz 2 StGB gesichert. Die Schweigepflicht ist hier mit dem Vertrauen in die amtliche Institution, statt, wie in Absatz 1, in die Angehörigen einer Berufsgruppe, begründet. Falls wichtige öffentliche Interessen gefährdet sind, können vorsätzliche oder fahrlässige Verstöße außerdem nach § 353b StGB [Verletzung des Dienstgeheimnisses und einer besonderen Geheimhaltungspflicht] verfolgt werden.

Die doppelte Verpflichtung dem einzelnen und der Gesellschaft gegenüber kann bei Psychologen wie bei Lehrern ***zu Konflikten*** führen, die eine ***Güterabwägung*** notwendig machen. Berufsethische Richtlinien können dabei Entscheidungshilfen sein; sie können und sie wollen dem Fachmann die Verantwortung nicht abnehmen. Ethische Regelwerke dürfen weder fachmethodische Entscheidungskalküle ersetzen, noch können sie ein darüber hinausreichendes Subsumtions-Raster für alle denkbaren Einzelfalle liefern.

11.2 Rechtsfragen

In der Bundesrepublik Deutschland gibt es nach wie vor kein verbindliches Berufsrecht für Psychologen. In der Praxis muß sich die Pädagogisch-psychologische Diagnostik zum einen an allgemeinen Gesetzesnormen orientieren, die auch diagnostisches Handeln betreffen, zum anderen an besonderen Rechtsvorschriften, die für das Arbeitsfeld “Schule” entsprechende Vorgaben enthalten.

In diesem Zusammenhang beschränken wir uns auf die praktisch wichtigsten Fragen der ***Zulässigkeit*** diagnostischer Maßnahmen sowie der ***rechtlichen Überprüfung*** der Maßnahmen und der Entscheidungen, die sich darauf stützen. Andere Aspekte, wie die Rechtsnatur diagnostischer Tätigkeit, Haftungsprobleme oder Spezialfragen der Eignungsdiagnostik und der Betätigung als Gerichtsgutachter, die nur einen losen Bezug zur Pädagogisch-psychologischen Diagnostik haben, bleiben hier ausgeklammert (s. dazu Kühne, 1987; Jessnitzer, 1988; Gaul, 1992; Zuschlag, 1992).

11.2.1 Zur Zulässigkeit Pädagogisch-psychologischer Diagnostik

Der rechtliche Rahmen, in dem sich die Pädagogisch-psychologische Diagnostik bewegen kann, ist in der Bundesrepublik durch das Grundgesetz (GG) von 1949 abgesteckt (s. Kasten auf dieser Seite). Art. 1 Abs. 1 GG erklärt die Menschenwürde für unantastbar und verpflichtet alle staatliche Gewalt, sie zu achten und zu schützen. Die in Art. 2 GG verbrieften ***allgemeinen Persönlichkeitsrechte*** auf freie Entfaltung und Unversehrtheit sind ein weiterer verfassungsrechtlich wichtiger Maßstab. In die Freiheitsrechte des Individuums darf nur auf der Grundlage von Gesetzen eingegriffen werden. Das allgemeine Persönlichkeitsrecht schützt die Privatsphäre des Bürgers vor Ausforschung. In Verbindung damit setzen auch andere Grundrechte der Pädagogisch-psychologischen Diagnostik Schranken. Vor dem Informationszugriff grundsätzlich geschützt sind Glaubensüberzeugungen (Art. 4 Abs. 1 GG) und die familiären Verhältnisse (Art. 6 Abs. 1 GG); ebensowenig sind Erhebungen zulässig, die der elterlichen Erziehungsverantwortung (Art. 6 Abs. 2 GG) zuwiderlaufen (Avenarius, 1990). Sollen zu Untersuchungszwecken Auskünfte erhoben werden, die die Privatsphäre oder den Intimbereich berühren, darf dies nur anonym auf freiwilliger Basis und bei Minderjährigen mit Einwilligung der Erziehungsberechtigten geschehen. Ohnehin bedarf es in der Regel der besonderen Genehmigung durch die Schulaufsichtsbehörde (s. weiter unten).

Verfassungsrechtliche Rahmenvorschriften
Auszug aus dem Grundrechtskatalog des Grundgesetzes für die Bundesrepublik Deutschland vom 23. Mai 1949

Art. 1. ***[Schutz der Menschenwürde]*** (1) Die Würde des Menschen ist unantastbar. Sie zu achten und zu schützen ist Verpflichtung aller staatlichen Gewalt.

(...)

Art. ***2. [Freiheitsrechte]*** (1) Jeder hat das Recht auf freie Entfaltung seiner Persönlichkeit, soweit er nicht die Rechte anderer verletzt und nicht gegen die verfassungsmäßige Ordnung oder das Sittengesetz verstößt.

(2) Jeder hat das Recht auf Leben und körperliche Unversehrtheit. Die Freiheit der Person ist unverletzlich. In diese Rechte darf nur auf Grund eines Gesetzes eingegriffen werden.

(...)

Art. 4. ***[Glaubens- und Bekenntnisfreiheit]*** (1) Die Freiheit des Glaubens, des Gewissens und die Freiheit des religiösen und weltanschaulichen Bekenntnisses sind unverletzlich.

(...)

Art. 6. ***[Ehe und Familie, nichteheliche Kinder]*** (1) Ehe und Familie stehen unter dem besonderen Schutze der staatlichen Ordnung.

(2) Pflege und Erziehung der Kinder sind das natürliche Recht der Eltern und die zuvörderst ihnen obliegende Pflicht. Über ihre Betätigung wacht die staatliche Gemeinschaft.

(...)

Art. 19. ***[Einschränkung von Grundrechten]***

(...)

(4) Wird jemand durch die öffentliche Gewalt in seinen Rechten verletzt, so steht ihm der Rechtsweg offen. Soweit eine andere Zuständigkeit nicht begründet ist, ist der ordentliche Rechtsweg gegeben. (...)

Rechtsvorschriften, die Schüler zur Teilnahme an diagnostischen Maßnahmen, insbesondere an Testverfahren, verpflichten, schränken das ***Recht auf informationelle Selbstbestimmung*** ein. Die Einschränkung muß daher dem rechsstaatlichen ***Gebot der Verhältnismäßigkeit*** genügen. Die Kriterien dafür sind die Geeignetheit, die Erforderlichkeit und die Zumutbarkeit der Maßnahme. Der Verwendungszweck der zu erhebenden Daten muß bereichsspezifisch und präzise bestimmt sein; die Daten dürfen nur im Rahmen dieser Zweckbindung verwendet werden (nach Avenarius, 1990, S. 33-34). Demzufolge darfallein das nach Maßgabe der Fragestellung notwendige Verfahren angewendet werden; dies auch nur, soweit es - neben seiner Eignung für den vorgesehenen Zweck - den Schülern zugemutet werden kann. Die Fragestellung ihrerseits muß sich selbstverständlich an den Rahmen des Zulässigen halten, der durch Gesetz oder Rechtsverordnung vorgegeben ist. Andere Daten mitzuerheben, ist nicht gestattet.

Uneingeschränkt zulässig sind offenbar nur die pflichtgemäß abzugebenden Lehrerurteile. Dies ist zwar logisch konsequent, empirisch wie rechtlich jedoch fragwürdig, weil Lehrerurteile ein subjektives Verfahren darstellen, das Verhalten und Leistungen von Schülern lediglich auf dem Niveau von Schätzskalen unklarer Güte wiedergibt. Ihre "Geeignetheit" für Fragestellungen von großer Tragweite, z.B. bei Einschulungs-, Umschulungs- oder Versetzungsentscheidungen, ist zweifelhaft, zumal ein breites Spektrum überprüfter objektiver Methoden zur Verfügung steht. Von daher verwundert es, daß sich die Frage der rechtlichen Zulässigkeit diagnostischer Maßnahmen einseitig auf die instrumentell meist besseren Testverfahren konzentriert.

In einigen Bundesländern können Lehrer und/oder Schulpsychologen auf der Grundlage von Bestimmungen in den Schulgesetzen und Rechtsverordnungen objektive Leistungsmessungen mit obligatorischer Teilnahme vornehmen. Bei der Feststellung der sog. Schulreife und der Entscheidung über die Sonderschuleinweisung sind die Schüler durchweg verpflichtet, sich einer diagnostischen Untersuchung zu unterziehen. Dies schließt die Anwendung von Testverfahren ein (Avenarius, 1990).

Anders liegen die Dinge im Falle der individuellen Schullaufbahn- und Bildungsberatung durch Beratungslehrer oder Schulpsychologen sowie der Beratung bei Lern- und Verhaltensstörungen durch Schulpsychologen. Sie wird zumeist auf eigenen Wunsch oder freiwillig in Anspruch genommen. Das Spektrum der zulässigen diagnostischen Verfahren ist hier weiter gefaßt und kann neben den Leistungs- auch psychometrische Persönlichkeitstests und projektive Methoden umfassen. Außer Schulleistungstests, die von Lehrern durchgeführt werden dürfen, ist die Anwendung Psychologen, teilweise auch Lehrern mit Zusatzausbildung (Beratungslehrer, Sonderschullehrer), vorbehalten (s. Kasten S. 237: Beispiel Hessen).

11.2.2 Zur rechtlichen Kontrolle diagnostischer Maßnahmen

Als Teil des staatlichen Verwaltungshandelns können diagnostische Maßnahmen von Aufsichtsbehörden innerhalb der Verwaltung (Exekutive) und von Instanzen der Rechtsprechung (Judikative) kontrolliert werden. Grundlegend für die rechtliche Beurteilung ist der in Art. 19 Abs. 4 GG garantierte Schutz des Bürgers vor Rechtsverletzung durch die öffentliche Gewalt (s. Kasten S. 235).

Beispiel Hessen

Tests und Erhebungen in Schulen; hier: Durchführung durch Lehrer. Erlaß vom 25.9.1985 (Amtsblatt des Hessischen Kultusministers (...), 38, 800-801)

I.

1. Zur Feststellung des Lernerfolges und von Lerndefiziten können in der Schule zwei Arten von Schulleistungstests durchgeführt werden:
a. Standardisierte Schulleistungstests (...)
b. Informelle Schulleistungstests (...)
Diese Tests dürfen von Lehrern durchgeführt werden, wenn ihnen Möglichkeiten und Grenzen der Testanwendung in der Schule allgemein bekannt sind und wenn sie die Methoden der Testdurchführung, -auswertung und -interpretation sicher beherrschen.

(...)

2. Die Durchführung anderer als der in Abs. 1 genannten Tests (z.B. Intelligenz- und Begabungstests) bedarf der Zustimmung der Erziehungsberechtigten oder der volljährigen Schüler. Sie sind auf die Freiwilligkeit der Angaben hinzuweisen. Aus einer Verweigerung von Angaben entstehen keine Rechtsnachteile. Der Schulelternbeirat ist zu unterrichten. Die Testergebnisse sind den Erziehungsberechtigten und den volljährigen Schülern auf Verlangen bekanntzugeben. Solche Tests sollen wegen der besonderen Schwierigkeiten bei der Durchführung und der Interpretation der Ergebnisse nur durch besonders ausgebildete Lehrer in Absprache mit dem Schulpsychologischen Dienst oder durch Schulpsychologen durchgeführt werden. Vorkehrungen zur Wahrung des Datenschutzes sind zu treffen.

(...)

IV

Schülerbefragungen mit verschiedenen Erhebungsmethoden (Fragebogen, Schätzskalen, Interviewtechniken u.ä.) und Datenerhebungen nach wissenschaftlichen Grundsätzen bedürfen der Zustimmung bzw. Anordnung des Staatlichen Schulamtes. Die Zustimmung ist nur dann zu erteilen, wenn wissenschaftliche Kriterien angemessen berücksichtigt werden, wenn sichergestellt ist, daß die Freiwilligkeit der Beteiligung und die Anonymität der Befragten gewahrt bleiben sowie deren Privatsphäre nicht berührt wird und wenn die Befragung bzw. Erhebung schulischen oder pädagogischen Zwecken dient und durch sie keine unangemessene Beeinträchtigung des Unterrichts erfolgt. Bei Befragung Minderjähriger müssen die Erziehungsberechtigten zustimmen. Absatz 1.2, Satz 2 und 3, gelten entsprechend. Andere Untersuchungen durch Lehrer oder Schülervertretungen zur Erforschung von Meinungen, Einstellungen und Werthaltungen bedürfen unter Beachtung der Bestimmungen der Allgemeinen Konferenzordnung der Zustimmung des Schulleiters.

V.

Die Durchführung von Verfahren, deren Anwendung ein abgeschlossenes Studium der Psychologie voraussetzen, sind Lehrern in der Schule nicht gestattet.

Zu diesen Verfahren gehören:
A) Leistungstests
   1. Intelligenztests (...)
   2. Allgemeine Leistungstests (...)
   3. Tests zur Prüfung spezieller Funktionen und Fähigkeiten (...)
B) Psychometrische Persönlichkeitstests
   1. Persönlichkeits-Struktur-Tests (...)
   2. Einstellungs- und Interessentests (...)
   3. Klinische Tests (...)
C) Projektive Verfahren (Entfaltungstests)
   1. Formdeuteverfahren (...)
   2. Verbalthematische Verfahren (...)
   3. Gestaltungs- und Wahlverfahren (...)

Abweichend davon dürfen Sonderschullehrer solche Testverfahren in der Schule durchführen, die sie nachweisbar während ihrer Ausbildung anzuwenden gelernt haben und für die eine Zustimmung der Erziehungsberechtigten vorliegt.
(...)

Die ***gerichtliche Überprüfung*** erstreckt sich in der Regel auf formale, für Schule und Diagnostik unspezifische Aspekte. Wegen fehlender gesetzlicher Regelung wird nach richterlichem Gewohnheitsrecht geprüft,
(a) ob von falschen Tatsachen ausgegangen wurde
(b) ob die geltenden Verfahrensvorschriften beachtet worden sind
(c) ob die Entscheidungsträger sich von sachfremden Einflüssen haben leiten lassen und
(d) ob allgemein anerkannte Bewertungsmaßstäbe beachtet wurden.

Eine rechtliche Kontrolle der inhaltlichen pädagogischen oder psychologischen Komponente findet in der Regel nicht statt (Berkemann, 1989). Das pädagogische oder psychologische Fachurteil gilt als juristisch nicht überprüfbar. Man räumt den Fachleuten einen Beurteilungsspielraum ein, der sich der rechtlichen Bewertung entziehe. Es wird überwiegend davon ausgegangen, daß es allgemein verbindliche, objektive Regeln und Kriterien der Leistungsbeurteilung nicht gibt.

Diese Argumentation wird zu Recht bezweifelt, u.a. mit dem Hinweis auf die Forschungslage und die Existenz praxisreifer Prototypen für ein diagnostisches Vorgehen, das sich auf einen breiten fachwissenschaftlichen Konsens stützen kann (Krapp, 1989). Welche Konsequenzen daraus gezogen werden sollen, erscheint offen. Krapp warnt vor voreiligen Entscheidungen. Er empfiehlt, das Instrumentarium der richterlichen Kontrolle vorsichtig und schrittweise zu verbessern. Dies sei möglich, weil die anerkannten Bewertungsgrundsätze, die für eine rechtliche Normierung benötigt werden, für Teilbereiche der Leistungsdiagnostik bereits vorlägen oder aufgestellt werden könnten. Dem können wir nur beipflichten.

Die ***administrative Kontrolle*** diagnostischer Maßnahmen ist in erster Linie eine Angelegenheit der ***Fachaufsicht.*** Im Unterschied zur Dienstaufsicht hat die Fachaufsicht das Verwaltungshandeln, einschließlich der Ermessenshandhabung, nach fach-

spezifischen Kriterien auf seine Sachgerechtheit und Zweckmäßigkeit zu prüfen. Neben der Würdigung von Einzelfallentscheidungen - wie in der Regel bei der gerichtlichen Kontrolle - geht es hier auch um die Beurteilung von Untersuchungsvorhaben, Verfahrensweisen und Methoden. Die Fachaufsicht wird von der vorgeordneten der nachgeordneten Behörde gegenüber ausgeübt und umfaßt außer der Informations- und Kontrollbefugnis auch die Weisungsbefugnis und das Recht zur Aufhebung von Entscheidungen.

Für die schulpsychologischen Dienste besteht insofern grundsätzliche Weisungsabhängigkeit, als ihnen untersagt oder auferlegt werden kann, bestimmte diagnostische Maßnahmen zu treffen oder bestimmte Methoden anzuwenden. Für die Wirksamkeit der Weisung ist dabei rechtlich unerheblich, ob der Weisungsbefugte über Fachkompetenz verfügt und die Weisung fachlichen Kriterien standhält. In einigen Bundesländern sind die Schulpsychologen jedoch im Hinblick auf die Datengewinnung und die Erstellung von Gutachten ausdrücklich weisungsfrei gestellt (vgl. Kühne, 1987, Teil 1 $ 5, Teil 11 § 2b).

Die Anwendung diagnostischer Methoden bei ***wissenschaftlichen Untersuchungen im Schulbereich*** unterliegt darüber hinaus generell der Kontrolle durch Schulaufsichtsbehörden. Bundesweit hat sich die Praxis durchgesetzt, die Durchführung empirischer Studien durch Externe nicht mehr der Vereinbarung vor Ort und damit dem Ermessen der Lehrer oder der Schulen zu überlassen. Die Vorhaben bedürfen der vorherigen Genehmigung durch Minister oder Regierungspräsidenten. Dies wird in erster Linie mit dem Schutz der Schulen vor unzumutbarer Belastung durch ein Übermaß an Untersuchungen begründet.

Die grundgesetzlich verbürgte ***Forschungsfreiheit*** (Art. 5 Abs. 3 GG) verpflichtet zwar den Staat zur Mitwirkung, indem er den Zugang zum Forschungsfeld eröffnet; die vom Forscher eingeforderte Kooperationsverpflichtung konkurriert jedoch mit der Verpflichtung, einen störungsfreien Schulbetrieb und Unterrichtsablauf zu gewährleisten (vgl. Avenarius, 1980). Die z.T. restriktive Handhabung von Anträgen hat allerdings den Eindruck aufkommen lassen, manche Schulbehörden wollten das Bildungswesen mit vorgeschobenen Gründen gegen Untersuchungen “abschotten” (Ingenkamp, 1980).

Die Genehmigung wird in der Regel davon abhängig gemacht, daß die Untersuchung wissenschaftlichen Ansprüchen genügt, keine unzumutbare Belastung für Schule, Schüler und Lehrer darstellt und, vor allem wenn sie während des Unterrichts durchgeführt werden soll, daß sie pädagogisch relevant ist. Die Anträge müssen neben der detaillierten Projektbeschreibung ***alle Erhebungsunterlagen*** (Tests, Fragebogen u.ä.) enthalten. Weitere übliche Auflagen betreffen die Wahrung von Anonymität und Freiwilligkeit, den Schutz der Intimsphäre und im Regelfall die Einwilligung der Erziehungsberechtigten (s. z.B. Hessischer Kultusminister, 1987). Das Verfahren ermöglicht die fachaufsichtliche Einwirkung auf das Projekt, hat eine bedenkliche forschungssteuernde Wirkung und macht repräsentative empirische Untersuchungen nahezu unmöglich. Für einen vernünftigen Ausgleich zwischen den konkurrierenden Rechtsgütern und eine gedeihliche Kooperation zwischen Wissenschaft und Schulbehörde empfiehlt Avenarius (1980), die Rahmenbedingungen für den Zugang zu den Schulen gesetzlich zu präzisieren und für das Genehmigungsverfahren Gutachterausschüsse einzurichten.

Zusammenfassung

Berufsethische Richtlinien sollen zum einen gewährleisten, daß die diagnostische Praxis, ihrer Bedeutung entsprechend, hohen professionellen Standards genügt, zum anderen können sie im Hinblick auf die doppelte Verantwortung gegenüber Individuum und Gesellschaft Orientierungshilfen bieten. Sie können dem Diagnostiker aber nicht die Verantwortung für sein Handeln abnehmen.

Dem diagnostischen Handeln sind aus dem Grundgesetz ableitbare rechtliche Schranken gesetzt. In die allgemeinen Persönlichkeitsrechte des Bürgers darf nicht unbefugt eingegriffen werden. Zu den geschützten Rechten gehört das Recht auf informationelle Selbstbestimmung. Diagnostische Maßnahmen und die darauf gestützten Entscheidungen können gerichtlich auf formale Mängel im Vorgehen überprüft werden. Ihre ***pädagogisch-psychologische Zweckmäßigkeit*** unterliegt der fachaufsichtlichen Kontrolle durch übergeordnete Behörden.

Diagnostische Erhebungen im Rahmen wissenschaftlicher Untersuchungen an Schulen sind genehmigungspflichtig. Die Schulen sollen vor unzumutbaren Belastungen bewahrt werden. Die Genehmigung kann mit Auflagen verbunden sein.

Weiterführende Literatur:

Blanke, Th. & Sterzel, D. (1991). Menschenwürde und Tests: Voraussetzungen und Grenzen ihrer rechtlichen Zulässigkeit. In S. Grubitzsch (Hrsg.), ***Testtheorie - Testpraxis. Psychologische Tests und Prüfverfahren im kritischen Überblick (S. 325-*** 372). Reinbek: Rowohlt.

Heckel, H. & Avenarius, H. (1986). ***Schulrechtskunde*** (6. Aufl.). Neuwied: Luchterhand.

Lecher, Th. (1988). ***Datenschutz und psychologische Forschung.*** Göttingen: Hogrefe.

Riegel, R. (1988). ***Datenschutz in der Bundesrepublik Deutschland.*** Heidelberg: von Decker & Müller.

Literaturverzeichnis

Abel, J. (1989). Profilanalysen in der Schulforschung. Zeitschriftfür Pädagogische Psychologie, *3, 27-34.*

Amelang, M. & Bartussek, D. (1990). *Differentielle Psychologie und Persönlichkeitsforschung* (3., überarbeitete u. erweiterte Aufl.). Stuttgart: Kohlhammer.

Amthauer, R. (1953). *Intelligenz-Struktur-Test.* Göttingen: Hogrefe.

Amthauer, R. (1970). *Intelligenz-Struktur-Test* (I-S-T 70) (4. Aufl., 1973). Göttingen: Hogrefe.

Anastasi, A. (1968). *Psychological Testing* (3rd ed.). New York: Macmillan.

Anderson, J.C. & Gerbing, D.W. (1988). Structural equation modeling in practice: A review and recommended two step approach. *Psychological Bulletin,* 103, 411-423.

Anderson, T.W. (1958). *An introduction to multivariate statistical analysis.* New York: Wiley.

Avenarius, H. (1980). Informationszugang - Forschungsfinanzierung - Publikationsfreiheit: Rechtsfragen im Verhältnis zwischen pädagogischer Forschung und Staat. In H. Avenarius, K. Ingenkamp & *G.* Otto, *Forschung und Lehre sind frei ... Wie die pädagogische Forschung von ihrem Gegenstand ausgesperrt wird* (S. 69-98). Weinheim: Beltz.

Avenarius, H. (1990). *Anwendung Diagnostischer Testverfahren in der Schule. Ein Rechtsgutachten.* Weinheim: Beltz.

Baumert, *I.* (1973). *Untersuchungen zur diagnostischen Valenz des HAWIK und die Entwicklung einer Kurzform (WIPKI).* Bern: Huber.

Belser, H. (1967). *Testentwicklung: Verfahren und Probleme der Entwicklung von Gruppen-Intelligenztests, dargestellt am Beispiel des Frankfurter Analogietests.* Weinheim: Beltz.

Bender, H. (1985). *Persönlichkeitstheorien von Grundschullehrern. Untersuchungen zu den impliziten Persönlichkeitstheorien von Lehrern in vierten Grundschulklassen.* Weinheim: Beltz.

Bennett, N. (1979). *Unterrichtsstil und Schülerleistung.* Stuttgart: Klett.

Bentler, PM. (1985). *Theory and implementation of EQS. A structural equations program.* Los Angeles. BMDP Statistical Software.

Bereiter, C. (1963). Some persisting dilemmas in the measurement of change. In C.W. Harris (Ed.), *Problems in measuring change.* Madison: University of Wisconsin Press.

Bergan, J.R. & Stone, C.A. (1985). Latent class models for knowledge domains. *Psychological Bulletin, 98,* 166-184.

Berkemann, J. (1989). Pädagogische Maßstäbe in der gerichtlichen Kontrolle schulischer Leistungen. *Zeitschriftfür Pädagogik, 35, 535-548.*

Bernstein, I.H. (1987). *Applied multivariate analysis.* Chapter 7: Confirmatory factor analysis (pp. 198-245). New York: Springer.

Berufsverband Deutscher Psychologen (1986). *Berufsordnung für Psychologen.* Bonn: Deutscher Psychologen Verlag.

Bessoth, R. (1983). *Lehrerbeurteilung.* Neuwied: Luchterhand.

Bierhoff, H.W. (1986). *Personenwahrnehmung. Vom ersten Eindruck zur sozialen Interaktion.* Berlin: Springer.

Birkel, P. (1978). *Mündliche Prüfungen.* Bochum: Kamp.

Blanke, Th. & Sterzel, D. (1991). Menschenwürde und Tests: Voraussetzungen und Grenzen ihrer rechtlichen Zulässigkeit. In S. Grubitzsch (Hrsg.), *Testtheorie - Testpraxis. Psychologische Tests und Prüfverfahren im kritischen Überblick* (S. 325-372). Reinbek: Rowohlt.

Blickle, G. (1991). Anregungsbedingungen für scheinbar paradox(al)e Wirkungen von Lob und Tadel. *Zeitschrift für Pädagogische Psychologie, 5, 21-32.*

Bloom, BS. (Hrsg.) (1972). *Taxonomie von Lernzielen im kognitiven Bereich* (amerikan. Original seit 1956). Weinheim: Beltz.

Bloom, B.S., Hastings, J.Th. & Madaus, G.F. (Eds.) (1971). *Handbook of formative and summative evaluation of student learning.* New York: McGraw Hill.

Bloxom, B. (1989). Adaptive Testing: A review of recent results. *Zeitschriftfür Differentielle und Diagnostische Psychologie,* 10, 1-17.

Borg, I. & Staufenbiel, Th. (1989). *Theorien und Methoden der Skalierung. Eine Einführung.* Bern: Huber.

Bortz, J. (1989). *Statistikfür Sozialwissenschaftler* (3. Aufl.). Berlin: Springer.

Brezinka, W. (1978). *Metatheorie der Erziehung. Eine Einführung in die Grundlagen der Erziehungswissenschaft, der Philosophie der Erziehung und der Praktischen Pädagogik (4.* Aufl.). München: Reinhardt.

Brezinka, W. (1981). *Erziehungsziele, Erziehungsmittel, Erziehungserfolg. Beiträge zu einem System der Erziehungswissenschaft* (2. Aufl.). München: Reinhardt.

Brezinka, W. (1986). Berufsethos des Lehrers: Ein vernachlässigtes Problem der Erziehungspolitik. In W. Brezinka, *Erziehung in einer wertunsicheren Gesellschaft (S.* 169-218). München: Reinhardt.

Bronfenbrenner, U. (1974). *Wie wirksam ist kompensatorische Erziehung?* Stuttgart: Klett.

Campbell, D.T. & Erlebacher, A.E. (1975). How regression artifacts in quasi-experimental evaluations can mistakenly make compensatory education look harmful. In E.L. Struening & M. Guttentag (Eds.), *Handbook of evaluation research* (Vol. 1, pp.597-617). Beverly Hills, California: Sage Publications.

Campbell, D.T. & Fiske, D.W. (1959). Convergent and discriminant Validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Campbell, D.T. & Stanley, J.L. (1963). Experimental and quasi-experimental designs for research on teaching. In N.L. Gage (Ed.), *Handbook of research on teaching* (pp.171-246). Chicago: Rand McNally.

Carlson, J.S. & Wiedl, K.H. (1976). Modes of presentation of the Raven Coloured Progressive Matrices Test: Toward a differential testing approach. *Trierer Psychologische Berichte,* 3, Heft 7.

Carlson, J.S. & Wiedl, K.H. (1980). Applications of a dynamic testing approach in intelligence assessment: Empirical results and theoretical formulations. *Zeitschriftfür Differentielle und Diagnostische Psychologie,* 1, 303-318.

Chow, S.L. (1990). Teachers' expectancy and its effects: A tutorial review. *Zeitschriftfür Pädagogische Psychologie, 4,* 147-159.

Cleary, T.A. (1968). Test bias: Prediction of grades of negro and white students in integrated Colleges. *Journal of Educational Measurement,* 5, 115-124.

Cohen, R. (1969). *Systematische Tendenzen bei Persönlichkeitsbeurteilungen.* Bern: Huber.

Cole, N.S. (1973). Bias in selection. *Journal of Educational Measurement,* 10, 237-255.

Comenius, A. (Komensky, J.A.) (dt. Ausg. 1959). Analytische Didaktik. Berlin: Volk und Wissen.

Conrad, W. (1983). Intelligenzdiagnostik. In K.-J. Groffmann & L. Michel (Hrsg.) *Intelligenz und Leistungsdiagnostik* (S. 104-201). Göttingen: Hogrefe.

Conrad, W., Baumann, E. & Mohr, V. (1980). *Mannheimer Test zur Erfassung des physikalisch-technischen Problemlösens.* Göttingen: Hogrefe.

Cook, Th. D. & Campbell, D.T. (1979). Quasi-Experimentation. *Design & analysis issues for field settings.* Chicago: Rand McNally College Publ. Comp.

Cronbach, L.J. (1970). *Essentials of psychological testing* (3rd ed.). New York: Harper & Row.

Cronbach, L.J. (1983). *Designing evaluations of educational and social programs* (2nd ed.). San Francisco: Jossey-Bass Publishers.

Cronbach, L.J. & Gleser, *G.C.* (1965). *Psychological tests and personnel decisions.* Urbana: University of Illinois Press.

Cronbach, L.J., Rajaratnam, N. & Gleser, G.C. (1963). Theory of generalizability: a liberalization of reliability theory. *British Journal of Statistical Psychology,* 16, 137-163.

Cronbach, L.J. & Snow, R.E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions.* New York: Irvington/Naiburg.
Dahl, *G*. (1972). *Reduzierter Wechsler-Intelligenztest.* Meisenheim/Glan: Hain.
Diederich, J. (1988). *Didaktisches Denken.* Weinheim: Juventa.
Dietrich, Th. (Hrsg.) (1982). *Diepädagogische Bewegung "Vom Kinde aus".* Bad Heilbrunn: Klinkhardt.
Dobrick, M. & Hofer, M. (1991). *Aktion und Reaktion. Die Beachtung des Schülers im Handeln des Lehrers.* Göttingen: Hogrefe.
Dumin, J.H. & Scandura, J.M. (1977). Algorithmic approach to assessing behavior potential: Comparison with item forms. In J.M. Scandura, *Problem solving* (pp. 347-363). New York: Academic Press.
Ebel, R.L. (1962). Content standard test scores. *Educational and Psychological Measurement,* 22, 15-22.
Eckes, T. & Roßbach, H. (1980). *Clusteranalysen.* Stuttgart: Kohlhammer.
Ehlers, B., Ehlers, Th. & Makus, H. (1978). *Marburger Verhaltensliste (MVL).* Göttingen: Hogrefe.
Einhorn, H.J. & Bass, A.R. (1971). Methodical considerations relevant to discrimination in employment testing. *Psychological Bulletin, 75,* 261-269.
Engelbrecht, W. (1975). Validierung einer Berufseignungs-Testbatterie und Verwendung der Ergebnisse für eine computerunterstützte Testbefundinterpretation. *Diagnostica,* 21, 3-24 und 97-106.
Engelbrecht, W. (1978). Weiterentwicklung der maschinellen Testbefundinterpretation zur EUB-Testbatterie. *Diagnostica, 24, 39-49.*
Eysenck, H.J. (1975). *Die Ungleichheit der Menschen.* München: List.
Fahrmeir, L. & Harnerle, A. (Hrsg.) (1984). *Multivariate statistische Verfahren.* Berlin: de Gruyter.
Feger, B. (1984). Die Generierung von Testitems zu Lehrtexten. *Diagnostica, 30, 24-46.*
Fischer, G.H. (1968). Kritik der Klassischen Testtheorie. In G.H. Fischer (Hrsg.), *Psychologische Testtheorie* (S.54-77). Bern: Huber.
Fischer, G.H. (1974). *Einführung in die Theorie psychologischer Tests.* Bern: Huber.
Fischer, G.H. (1983). Neuere Testtheorie. In H. Feger & J. Bredenkamp (Hrsg.), *Messen und Testen* (S. 604-692). Göttingen: Hogrefe.
Fischer, G.H. & Roppert, J. (1964). *Bemerkungen zu einem Verfahren der Transformationsanalyse, Archivfür die gesamte Psychologie,* 116, 98-100.
Fiske, S.T. & Taylor, S.E. (1991). *Social Cognition.* New York: McGraw-Hill.
Flammer, A. (1974). Längsschnittuntersuchung mit Lern- und Transfertests. *Schweizerische Zeitschriftfür Psychologie, 33,* 14-32.
Flammer, A. (1975). *Individuelle Unterschiede im Lernen.* Weinheim: Beltz.
Flammer, A. & Schmid, H. (1982). Lerntests: Konzept, Realisierungen, Bewährung. Eine Übersicht. *Schweizerische Zeitschrift für Psychologie,* 41, 114-138.
Formann, A.K. (1984). *Die Latent-Class-Analyse.* Weinheim: Beltz.
Formann, A.K., Ehlers, Th. & Scheiblechner, H. (1980). Anwendung der Latent-Class-Analyse auf Probleme der diagnostischen Klassifikation am Beispiel der Marburger Verhaltensliste. *Zeitschriftfür Differentielle und Diagnostische Psychologie,* 1, 319-330.
Franzen, U. & Merz, F. (1976). Einfluß des Verbalisierens auf die Leistung bei Intelligenzprüfungen. Neue Untersuchungen. *Zeitschriftfür Entwicklungspsychologie und Pädagogische Psychologie,* 8, 117-134.
Fricke, R. (1972). Testgütekriterien bei lehrzielorientierten Tests. *Zeitschrift für erziehungswissenschaftliche Forschung, 6,* 150-175.
Fricke, R. (1974). *Kriteriumsorientierte Leistungsmessung.* Stuttgart: Kohlhammer.
Gaul, D. (1992). *Rechtsprobleme psychologischer Eignungsdiagnostik.* Bonn: Deutscher Psychologen Verlag.

Gittler, G. (1984). Entwicklung und Erprobung eines neuen Testinstruments zur Messung des räumlichen Vorstellungsvermögens. *Zeitschrift für Differentielle und Diagnostische Psychologie,* 5, 141-165.

Gittler, G. & Wild, B. (1988). Der Einsatz des LLTM bei der Konstruktion eines Itempools für das adaptive Testen. In K.D. Kubinger (Hrsg.), *Moderne Testtheorie - Ein Abriß samt neuesten Beitrügen* (S. 115-139). Weinheim: Psychologie Verlags Union.

Glas, C.A.W. *(1989). Contributions to estimating and testing Rasch-Models.* Dissertation der Universität Twente. Den Haag: CIP-Gegevens Koninklijke Bibliotheek (ISBN 90-900 30786).

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist,* 18, 519-521.

Glaser, R. (1973). Unterrichtstechnologie und die Messung von Lernergebnissen: Einige Fragen. In P. Strittmatter (Hrsg.), *Lernzielorientierte Leistungsmessung.* Weinheim: Beltz.

Glöckel, H. (1992). *Vom Unterricht. Lehrbuch der Allgemeinen Didaktik* (2. Aufl.). Bad Heilbrunn: Klinkhardt.

Goldman, St.H. & Raju, N.S. (1986). Recovery of one- and-two-parameter logistic item parameters: An empirical study. *Educational and Psychological Measurement, 46,* 11-35.

Guba, E.G. & Lincoln, *Y.S. (1982). Effective Evaluation* (2nd ed.). San Francisco: Jossey-Bass Publishers.

Günther, Ch. & Günther, R. (1981). Zur Bedingungsanalyse von Intelligenzleistungen Erwachsener - Eine Untersuchung mit einem Langzeitlerntest. *Zeitschrift für Psychologie,* 189, 407-421.

Guthke, J. (1972). *Zur Diagnostik der intellektuellen Lernfähigkeit.* Berlin: Deutscher Verlag der Wissenschaften (Stuttgart: Klett, 1977).

Guthke, J. (1976). Entwicklungsstand und Probleme der Lernfähigkeitsdiagnostik. Teil I und II. *Zeitschriftfür Psychologie,* 184, 103-117 und 215-239.

Guthke, J. (1980a). *Ist Intelligenz meßbar?* Berlin: Deutscher Verlag der Wissenschaften.

Guthke, J. (1980b). Die Relevanz des Lerntestkonzepts für die klinisch-psychologische Diagnostik - demonstriert am Beispiel der geistigen Behinderung und der Frühkindlichen Hirn-Schädigung. *Probleme und Ergebnisse der Psychologie, 72,* 5-21.

Guthke, J. (1982). The learning test concept - an alternative to the traditional static intelligen*ce* test. *The German Journal of Psychology, 6, 306-324.*

Guthke, J. & Lehwald, G. (1984). On component analysis of the intellectual learning ability in learning tests. *Zeitschriftfür Psychologie,* 192, 3-17.

Guthke, J., Räder, E., Caruso, M. & Schmidt, K.D. (1991). Entwicklung eines adaptiven computergestützten Lerntests auf der Basis der strukturellen Informationstheorie. *Diagnostica,* 37, 1-28.

Haertel, E.H. (1990). *Continuous and discrete latent structure models for item response data. Psychometrika, 55, 477-494.*

Hager, W. & Weißmann, *S.* (1991). *Bestätigungstendenzen in der Urteilsbildung.* Göttingen: Hogrefe.

Hambleton, R.K. & Cook, L.L. (1983). Robustness of item response models and effect of test lenght and sample size on the precision of ability estimates. In D.J. Weiss (Ed.), *New horizons in testing* (pp.31-49). New York: Academic Press.

Hardesty, A. & Lauber, H. (1956). *Hamburg-Wechsler-Intelligenztest für Erwachsene (HAWIE).* Bern: Huber.

Harris, C.W. (Hrsg.) (1963). *Problems in measuring change.* Madison: University of Wisconsin Press.

Hartig, M. (1975). *Probleme und Methoden der Psychotherapieforschung.* München: Urban & Schwarzenberg.

Hartmann, H.A. & Haubl, R. (Hrsg.) (1984). *Psychologische Begutachtung. Problembereiche und Praxisfelder.* München: Urban & Schwarzenberg.

Hartung, J. & Elpelt, B. (1984). *Multivariate Statistik.* München: Oldenbourg.
Hasemann, K. (1983). Verhaltensbeobachtung und Ratingverfahren. In K.-J. Groffmann & L. Michel (Hrsg.), *Verhaltensdiagnostik* (S. 434-488). Göttingen: Hogrefe.
Heckel, H. & Avenarius, H. (1986). *Schulrechtskunde* (6. Aufl.). Neuwied: Luchterhand.
Herbig, M. (1973). Beurteilung lehrzielorientierter Tests entsprechend den Methoden der Klassischen Testtheorie? In P. Strittmatter (Hrsg.), *Lernzielorientierte Leistungsmessung (S.* 156-178). Weinheim: Beltz.
Herrmann, Th. (1973). *Persönlichkeitsmerkmale. Bestimmung und Verwendung in der psychologischen Wissenschuft.* Stuttgart: Kohlhammer.
Herrmann, Th. (1991). *Lehrbuch der empirischen Persönlichkeitsforschung (6.* Aufl.). Göttingen: Hogrefe.
Hessischer Kultusminister (1985). Tests und Erhebungen in Schulen. Erlaß vom 25.9.1985. *Amtsblatt des Hessischen Kultusministers und des Hessischen Ministers für Wissenschaft und Kunst, 38, 800-801.*
Hessischer Kultusminister (1987). Wissenschaftliche Untersuchungen im Schulbereich. Erlaß vom *29.9.1987. Amtsblatt des Hessischen Kultusministers und des Hessischen Ministers für Wissenschaft und Kunst, 40, 756.*
Hilke, R. (1980). *Grundlagen* normorientierter *und* kriteriumorientierter Tests. Bern: Huber.
Hively, W., Patterson, H.L. & Page, S.H. (1968). A "universe-defined" system of arithmetic 'achievement tests. *Journal of Educational Measurement, 5,* 275-290.
Hofer, M. (1986). *Sozialpsychologie erzieherischen Handelns. Wie das Denken und Verhalten von Lehrern organisiert ist.* Göttingen: Hogrefe.
Hofer, M. & Pikowsky, B. (1988). Wie Jugendliche bei freier Antwortmöglichkeit Lehrersanktionen deuten. *Zeitschriftfür Pädagogische Psychologie, 2, 243-250.*
Hofstätter, P.R. (1977). *Persönlichkeitsforschung* (2. Aufl.). Stuttgart: Kröner.
Horn, W. (1983). *Leistungsprüfsystem (LPS).* (2. Aufl.). Göttingen: Hogrefe.
Hornke, L.F. & Habon, M.W. (1984). Erfahrungen zur rationalen Konstruktion von Testaufgaben. *Zeitschriftfür Differentielle und Diagnostische Psychologie, 5, 203-212.*
Huber, H.P. (1973). *Psychometrische Einzelfalldiagnostik.* Weinheim: Beltz.
Hussy, W. & Wiedl, K.H. (1978). Initialverhalten und "Restlösungsmenge" bei verschiedenen Lerntestprozeduren im Farbigen Matrizentest. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 10,* 157-168.
Ingenkamp, K. (1975). *Pädagogische Diagnostik. Ein Forschungsbericht über Schülerbeurteilung in Europa. Trendbericht im Auftrag des Europarats Straßburg.* Weinheim: Beltz.
Ingenkamp, K. (1980). Die Freiheit der pädagogisch-psychologischen Forschung - Verfassungsauftrag und Realität. In H. Avenarius, K. Ingenkamp & *G.* Otto, *Forschung und Lehre sind frei . . . Wie die pädagogische Forschung von ihrem Gegenstand ausgesperrt wird* (S. 45-67). Weinheim: Beltz.
Ingenkamp, K. (1985). *Lehrbuch der Pädagogischen Diagnostik* (Studienausgabe, 1988). Weinheim: Beltz.
Ingenkamp, K. (1989). Testkritik ohne Alternative. Eine kritische Darstellung der Argumentation radikaler Schultestkritik in der deutschen Fachliteratur. In K. Ingenkamp, *Diagnostik in der Schule* (S. 207-249). Weinheim: Beltz.
Ingenkamp, K. (1990). *Pädagogische Diagnostik in Deutschland* 1885-1932. Weinheim: Deutscher Studien Verlag.
Irle, M. & Allehoff, W. (1984). *Berufs-Interessen-Test II (BIT II).* Hogrefe: Göttingen.
Jäger, R.S. (1986). *Der diagnostische Prozeß. Eine Diskussion psychologischer und methodischer Randbedingungen* (2., verbesserte Aufl.). Göttingen: Hogrefe.
Jäger, R.S. (Hrsg.) (1988). *Psychologische Diagnostik. Ein Lehrbuch.* München: Psychologie Verlags Union. [2. Aufl.: Jäger, R.S. & Petermann, F. (Hrsg.) (1992)].
Jessnitzer, K. (1988). *Der gerichtliche Sachverständige. Ein Handbuch für die Praxis auf wissenschaftlicher Grundlage* (9. Aufl.). Köln: Heymanns.

Jöreskog, K.G. (1967). Some contributions to maximum likelihood factor analysis. *Psychometrika, 32, 443-482.*

Jöreskog, K.G. (1969). Ageneral approach to confirmatory factor analysis. *Psychometrika, 34,* 183-202.

Jöreskog, K.G. (1978). Structural analysis of covariance and correlation matrices. *Psychometrika, 43, 443-477.*

Jöreskog, K.G. & Sörbom, D. *(1985). LISREL VI. User's guide.* Mooresville: Scientific Software, Inc.

Jöreskog, K.G. & Sörbom, D. (1989). *LISREL 7. A guide to theprogram and applications.* 2nd ed. Chicago: SPSS Inc.

Kallina, H. (1967). Das Unbehagen in der Faktorenanalyse *Psychologische Beiträge, 10, 81-86.*

Kallus, K.W. & Janke, W. (1988). Klassenzuordnung. In R.S. Jäger (Hrsg.), *Psychologische Diagnostik* (S.131-145). München: Psychologie Verlags Union.

Kalveram, K. (1965). Die Veränderung von Faktorenstrukturen durch "simultane Überlagerung". *Archiv für die gesamte Psychologie, 117, 296-305.*

Kalveram, K. (1969). Kornpensatorische Kovarianz als Beispiel für einen Selektionseffekt oder Wie man aus positiven Korrelationskoeffizienten negative macht. *Archiv für die gesamte Psychologie, 121, 255-265.*

Kalveram, K. (1970a). Über Faktorenanalyse. Kritik eines theoretischen Konzepts und seine mathematische Neuformulierung. *Archiv für Psychologie, 122,* 92-118.

Kalveram, K. (1970b). Probleme der Selektion in der Faktorenanalyse. *Archiv für Psychologie,* 122, Teil I 199-214, Teil II 215-222.

Kaminski, G. (1982). Rahmentheoretische Überlegungen zur Taxonomie psychodiagnostischer Prozesse. In K. Pawlik (Hrsg.), *Diagnose der Diagnostik. Beiträge zur Diskussion der psychologischen Diagnostik in der Verhaltensmodifikation* (2. Aufl., S. 45-70). Stuttgart: Klett.

Kellaghan, T.K., Madaus, G.F. & Airasian, P.W. (1982). *The effects of standardized testing.* Boston: Kluwer.

Kenny, D. (1976). An empirical application of confirmatory factor analysis to the multitrait-multimethod matrix. *Journal of Experimental Social Psychology, 12, 247-252.*

Kisser, R. (1988). Adaptive Strategien. In R.S. Jäger (Hrsg.), *Psychologische Diagnostik* (S. 123-130). München: Psychologie Verlags Union.

Klafki, W. (1991). Neue *Studien zur Bildungstheorie und Didaktik* (2. Aufl.). Weinheim: Beltz.

Klauer, K.J. (1974). *Methodik der Lehrzieldefinition und Lehrstoffanalyse.* Düsseldorf: Schwann.

Klauer, K.J. (Hrsg.) (1978). *Handbuch der Pädagogischen Diagnostik, Bünde 1 bis 4* (Taschenbuchausgabe in 2 Bänden, 1982). Düsseldorf: Schwarm.

Klauer, K.J. (1978). Kontentvalidität. In K.J. Klauer (Hrsg.), *Handbuch der Pädagogischen Diagnostik, Band* 1 (S. 225-255). Düsseldorf: Schwann.

Klauer, K.J. (1983). Kriteriumsorientierte Tests. In H. Feger & J. Bredenkamp (Hrsg.), *Messen und Testen* (S. 693-726). Göttingen: Hogrefe.

Klauer, K.J. (1987). *Kriteriumsorientierte Tests.* Göttingen: Hogrefe.

Kleber, E.W. (1992). *Diagnostik in pädagogischen Handlungsfeldern.* Weinheim: Juventa.

Kleiter, E. (1973). Über Theorie und Modell kategorialer Fehler des Lehrerurteils. *Psychologische Beitrüge, 15,* 185-229.

Kordes, H. (1984). Pädagogische Aktionsforschung. In H. Haft & H. Kordes (Hrsg.), Methoden der Erziehungs- und Bildungsforschung. *Enzyklopädie der Erziehungswissenschaft, Bd.* 2 (S. 185-219). Stuttgart: Klett-Kotta.

Kormann, A. (1979). Lerntests als Alternative zu herkömmlichen Statustests. In D. Bolscho & Ch. Schwarzer (Hrsg.), *Beurteilen in der Grundschule* (S. 146-161). München: Urban & Schwarzenberg.

Kotmann, A. (1982). Möglichkeiten von Lerntests für Diagnose und Optimierung von Lernprozessen. In H. Ingenkamp, R. Horn & R. Jäger (Hrsg.), *Tests und Trends (S.* 97-117). Weinheim, Beltz.

Kormann, A. & Sporer, S.L. (1983). Learning-Tests - Concepts and critical evaluation. *Studies in Educational Evaluation, 9,* 169-184.

Krampen, G. & von Delius, A. (1981). Zur direkten Messung subjektiv erlebter gesundheitlicher Veränderungen. *Medizinische Psychologie, 7,* 166-174.

Krapp, A. (1986). Diagnose und Prognose. In B. Weidenmann, A. Krapp, M. Hofer, G.L. Huber & H. Mandl (Hrsg.), *Pädagogische Psychologie* (S. 565-630). München: Psychologie Verlags Union.

Krapp, A. (1989). Der zweifelhafte Beitrag der empirischen Pädagogik zur rechtlichen Kontrolle der schulischen Leistungsbeurteilung. *Zeitschriftfür Pädagogik, 35, 549-564.*

Krauth, J. (1983a). Diskriminanzanalyse. In J. Bredenkamp & H. Feger (Hrsg.), *Strukturierung und Reduzierung von Daten* (S. 293-350). Göttingen: Hogrefe.

Krauth, J. (1983b). Typenanalyse. In J. Bredenkamp & H. Feger (Hrsg.), *Strukturierung und Reduzierung von Daten* (S. 440-496). Göttingen: Hogrefe.

Krauth, J. (1983c). Bewertung der Änderungssensitivität von Items. *Zeitschriftfür Differentielle und Diagnostische Psychologie, 4, 7-28.*

Krauth, J. (1983d). Methodische Probleme in der pädagogischen Evaluationsforschung. *Zeitschrift für Empirische Pädagogik, 7,* 1-21.

Kristof, W. (1958). Statistische Prüfverfahren zur Beurteilung von Testprofilen. *Zeitschriftfür experimentelle und angewandte Psychologie, 5, 520-533.*

Kriz, J. & Lisch, R. (1988). *Methoden-Lexikon für Mediziner, Psychologen, Soziologen.* München: Psychologie Verlags Union.

Kubinger, K.D. (1986). Adaptive Intelligenzdiagnostik. *Diagnostica, 32, 330-344.*

Kubinger, K.D. (1988). Aktueller Stand und kritische Würdigung der probabilistischen Testtheorie. In K.D. Kubinger (Hrsg.), *Moderne Testtheorie-Ein Abriß samt neuesten Beirrügen* (S. 19-83). Weinheim: Beltz.

Kubinger, K.D. & Wurst, E. (1985). *Adaptives Intelligenz Diagnostikum AID.* Manual. Weinheim: Beltz.

Kühn, R. (1983). *Bedingungen für Schulerfolg. Zusammenhänge zwischen Schülermerkmalen, häuslicher Umwelt und Schulnoten.* Göttingen: Hogrefe.

Kühne, H.-H. (Hrsg.) (1987). *Berufsrechtfür Psychologen.* Baden-Baden: Nomos.

Langeheine, R. & Van de Pol, E (1990). *Discrete time mixed Markov latent class models.* Netherlands Central Bureau of Statistics. P.O.Box 959, 2270 AZ Voorburg, The Netherlands.

Langeheine, R. & Rost, J. (Eds.) (1988). *Latent* trait *and latent class models. New* York: Plenum Press.

Langfeldt, H.-P. & Fingerhut, W. (1984). Empirische Ansätze zurAufklärung des Konstruktes "Schulleistung". In K.A. Heller (Hrsg.), *Leistungsdiagnostik in der Schule* (4., völlig neubearbeitete Aufl., S. 40-45). Bern: Huber.

Lay, W.A. (1903). *Experimentelle Pädagogik.* Wiesbaden: Nemnich.

Laux, H. (1990). *Pädagogische Psychologie im Nationalsozialismus* 1933-1945. Weinheim: Deutscher Studien Verlag.

Lazarsfeld, P.F. (1950). Logical and mathematical foundations of latent structure analysis. In S.A. Stouffer, L. Guttman, E.A. Suchman, PF. Lazarsfeld, S.A. Star, J.A. Clausen (Eds.), *Studies in Social psychology in world war II* (Vol. IV, pp. 362-412). Princeton: Princeton University Press.

Lazarsfeld, PF. & Henry, N.W. (1968). *Latent structure analysis.* Boston: Houghton Mifflin.

Lecher, Th. (1988). *Datenschutz und psychologische Forschung.* Göttingen: Hogrefe.

Legler, R. (1977). Ein Beitrag zur Lernfähigkeitsdiagnostik - Der PLT (Problemlösungslemtest). *Probleme und Ergebnisse der Psychologie, 63, 55-77.*

Lienert, G.A. (1961). *Testaufbau und Testanalyse.* Weinheim: Beltz.
Lienert, G.A. (1991). *Testaufbau und Testanalyse* (5. Aufl.). Weinheim: Psychologie Verlags Union.
Linn, R.L. (1973). Fair test use in selection. *Review of Educational Research,* 43, 139-161.
Lord, EM. (1964). Nominally and rigorously parallel test forms. *Psychomerrika, 29. 335-346.*
Lord, EM. & Novick, M.R. (1968). *Statistical theories of mental test scores.* Reading, Mass.: Addison-Wesley.
Ludwig, P. (1991). *Sich selbst erfüllende Prophezeiungen im Alltagsleben.* Stuttgart: Verlag für Angewandte Psychologie.
Mager, R.F. (1965). *Lernziele und programmierter Unterricht.* Weinheim: Beltz.
Majcen, A.M., Steyer, R. & Schwenkmezger, P. (1988). Konsistenz und Spezifität bei Eigenschafts- und Zustandsangst. *Zeitschrift für Differentielle und Diagnostische Psychologie, 9,* 105-120.
McBridge, J.R. & Martin, J.T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D.J. Weiss (Ed.), New *horizons in testing* (pp. 223-236). New York: Academic Press.
McDonald, R.P. (1985). *Factoranalysis and related methods.* Hillsdale: Erlbaum Ass.
Meili, R. & Steingrüber, H.-J. (1978). *Lehrbuch der psychologischen Diagnostik (6.* Aufl.). Bern: Huber.
Melchinger, H. (1981). Lerntests statt Intelligenztests? Ein empirischer Beitrag zum Lerntestkonzept. In K.J. Klauer & H.-J. Kornadt (Hrsg.), *Jahrbuch für empirische Erziehungswissenschaft* (S. 125-157). Düsseldorf: Schwann.
Merz, F. (1963). Die Beurteilung unserer Mitmenschen als Leistung. In G.A. Lienert (Hrsg.), *Bericht über den 32. Kongreß der Deutschen Gesellschaftfür Psychologie (S.* 32-51). Göttingen: Hogrefe.
Merz, F. & Kalveram, K. (1965). Kritik der Differenzierungshypothese der Intelligenz. *Archiv für die gesamte Psychologie, 117, 287-295.*
Merz, F. & Stelzl, I. (1973). Modellvorstellungen über die Entwicklung der Intelligenz in Kindheit und Jugend. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 5, 153-166.*
Merz, F. & Stelzl, *I.* (1977). *Einführung in die Erbpsychologie.* Stuttgart: Kohlhammer.
Meumann, E. (1907). *Vorlesungen zur Einführung in die experimentelle Pädagogik und ihre psychologischen Grundlagen, Bde. 1 und* 2. Leipzig: Engelmann.
Meyer, W.-U. (1984). *Das Konzept von der eigenen Begabung.* Bern: Huber.
Meyer, W.-U., Bedau, U. & Engler, N. (1988). Indirekte Mitteilungen über Fähigkeitseinschätzungen in hypothetischen Lehrer-Schüler-Interaktionen. *Zeitschriftfür Pädagogische Psychologie, 2, 235-242.*
Michel, L. & Conrad, W. (1982). Theoretische Grundlagen psychometrischer Tests. In K.-J. Groffmann & L. Michel (Hrsg.), *Grundlagen psychologischer Diagnostik (S.* 1-129). Göttingen: Hogrefe.
Michel, L. & Mai, N. (1969). Zur varianzanalytischen Schätzung der Auswertungsobjektivität und eine empirische Untersuchung des Hamburg-Wechsler-Intelligenz-Tests für Erwachsene (HAWIE). *Psychologische Beiträge, 11, 23-33.*
Millman, J. & Darling-Hammond, L. (Eds.) (1990). *The new handbook of teacher evaluation. Assessing elementary and secondary school teachers.* Newbury Park: Sage.
Möbus, C. (1978). Zur Fairness psychologischer Intelligenztests. Ein unlösbares Problem zwischen Gruppen, Individuen, Institutionen? *Diagnostica, 24,* 191-234.
Möbus, C. (1983). Die praktische Bedeutung der Testfairness als zusätzliches Kriterium zu Reliabilität und Validität. In R. Horn, K. Ingenkamp & R.S. Jäger (Hrsg.), *Tests und Trends* 3 (S. 155-203). Weinheim: Beltz.
Möbus, C. & Nagl, W. (1983). Messung, Analyse und Prognose von Veränderungen. In J. Bredenkamp & H. Feger (Hrsg.), *Hypothesenprüfung. Enzyklopädie der Psychologie, Serie I, Forschungsmethoden der Psychologie, Bd.* 5 (S. 239-470). Göttingen: Hogrefe.

Mogel, H. (1990). *Umwelt* und *Persönlichkeit. Bausteine einerpsychologischen Umwelttheorie.* Göttingen: Hogrefe.

Möller, Ch. (Hrsg.) (1974). *Praxis der Lernplanung.* Weinheim: Beltz.

Möller, Ch. *(1976). Technik der Lernplanung. Methoden und Probleme der Lernzielerstellung* (5. Aufl.). Weinheim: Beltz.

Murray, H.A. (1938). *Explorations in personality.* New York: Oxford University Press.

Nährer, W. (1980). Zur Analyse von Matrizenaufgaben mit dem linearen logistischen Testmodell. *Zeitschrift für experimentelle und angewandte Psychologie, 27, 553-564.*

Nährer, W. (1988). "Schnelligkeitsangepaßtes Testen": Testökonomie unter Berücksichtigung der Testzeit. In K.D. Kubinger (Hrsg.), *Moderne Testtheorie - Ein Abriß samt neuesten Beitrügen* (S. 219-236). Weinheim: Psychologie Verlags Union.

Noack, H. & Petermann, F. (1988). Entscheidungstheorie. In R.S. Jäger (Hrsg.), *Psychologische Diagnostik. Ein Lehrbuch* (S. 241-253). München: Psychologie Verlags Union.

Norden, I. (1930). Neubearbeitung der Binet-Methode. *Zeitschrift für Kinderforschung, 37,* 75-92.

Nußbaum, A. (1987). Das Modell der Generalisierbarkeitstheorie. In K.J. Klauer (Hrsg.), *Kriteriumsorientierte Tests* (S. 114-136). Göttingen: Hogrefe.

Oldenbürger, H.A. (1983). Clusteranalyse. In J. Bredenkamp & H. Feger (Hrsg.), *Strukturierung und Reduzierung von Daten (S. 390-439).* Göttingen: Hogrefe.

Osbum, H.G. (1968). Item-sampling for achievement testing. *Educational und Psychological Measurement, 28, 95-104.*

Ostendorf, F., Angleitner, A. & Ruch, W. (1986). *Die Multitrait-Multimethod Analyse. Konvergente und diskriminante Validität der Personality Research Form.* Göttingen: Hogrefe.

Oswald, W.D. & Roth, E. (1978). *Der Zahlen-Verbindungs-Test (ZVT).* Göttingen: Hogrefe.

Patton, M.Q. (1981). *Creative Evaluation.* London: Sage.

Pawlik, K. (1971). *Dimensionen des Verhaltens* (2. Aufl.). Bern: Huber.

Pawlik, K. (1982). Modell- und Praxisdimensionen psychologischer Diagnostik. In K. Pawlik (Hrsg.), *Diagnose der Diagnostik. Beiträge zur Diskussion der psychologischen Diagnostik in der Verhaltensmodifkation* (2. Aufl., S. 13-43). Stuttgart: Klett.

Petermann, F. (1986). Probleme und neuere Entwicklungen der Veränderungsmessung - ein Überblick. *Diagnostica, 32, 4-16.*

Popham, W.J. & Husek, T.R. (1973). Implikationen kriteriumsbezogener Messungen. In P. Strittmatter (Hrsg.), *Lernzielorientierte Leistungsmessung* (S.46-58). Weinheim: Beltz.

Preiser, *S.* (1979). *Personwahrnehmung und Beurteilung.* Darmstadt: Wissenschaftl. Buchgesellschaft.

Rajaratnam, N., Cronbach, L.J. & Gleser, G.C. (1965). Generalizability of stratified parallel tests. *Psychometrika, 30. 39-56.*

Rauchfleisch, U. (1979). *Handbuch zum Rosenzweig Picture Frustration Test (PFT). 2* Bde. Bern: Huber.

Raven, J.R. (1963). *Guide to using the Coloured Progressive Matrices.* London: Lewis.

Revenstorf, D. (1980). *Faktorenanalyse.* Stuttgart: Kohlhammer.

Rheinberg, F. (1980). *Leistungsbewertung und Lernmotivation.* Göttingen: Hogrefe.

Rheinberg, F. & Weich, K.-W. (1988). Wie gefährlich ist Lob? Eine Untersuchung zum "paradoxen Effekt" von Lehrersanktionen, *Zeitschriftfür Pädagogische Psychologie, 2, 227-233.*

Riegel, R. (1988). *Datenschurz in der Bundesrepublik Deutschland.* Heidelberg: von Decker & Müller.

Rindskopf, D. (1983). A general framework for using latent class analysis to test hierarchical and nonhierarchical learning models. *Psychometrika, 48, 85-97.*

Roid, G.H. & Haladyna, Th. M. (1982). *A technology for test-item writing.* New York: Academic Press.

Rollett, B. (1978). Gruppierung von Schülern. In K.J. Klauer (Hrsg.), *Handbuch der Pädagogischen Diagnostik, Bd.* 4 (S. 893-915). Düsseldorf: Schwann.

Rollett, B. (1985). Testbesprechung über den "Mengenfolgen-Test. Kurzzeitlerntest für Schulanfänger (MIT) von J. Guthke". *Zeitschrift für Differentielle und Diagnostische Psychologie, 6,* 185-187.

Rollett, B. & Bartram, M. (1977). *Anstrengungsvermeidungstest.* Braunschweig: Westermann.

Rosenthal, R. (1975). Der Pygmalion-Effekt lebt. *Psychologie heute, 2 (Heft 61,* 18-21, 76-77.

Rosenthal, R. & Jacobson, L. (1971). *Pygmalion im Unterricht. Lehrererwartungen und Intelligenzentwicklung der Schüler.* Weinheim: Beltz (Original 1968: Pygmalion in the Classroom. New York: Holt).

Rost, D.H. (1987). Leseverständnis oder Leseverständnisse? *Zeitschriftfür Pädagogische Psychologie, 1,* 175-196.

Rost, J. (1977). *Diagnostik des Lernzuwachses. IPN-Arbeitsbericht Nr.* 26. Kiel: Institut für die Pädagogik der Naturwissenschaften, Olshausenstraße 40-60.

Rost, J. (1988). *Quantitative und qualitative probabilistische Testtheorie.* Bern: Huber.

Rost, J. & Spada, H. (1983). Die Quantifizierung von Lerneffekten anhand von Testdaten. *Zeitschriftfür Differentielle und Diagnostische Psychologie, 4, 29-49.*

Roth, L. (Hrsg.) (1991). *Pädagogik. Handbuchfür Studium und Praxis.* München: Ehrenwirth.

Sauer, K. (1981). *Einführung in die Theorie der Schule.* Darmstadt: Wissenschaftl. Buchgesellschaft.

Scandura, J.M. (1977). *Problem solving.* With co-contributions by J. Durnin, W. Ehrenpreis, G. Lowerre, W. Reulecke, D. Voorhies & W. Wulfeck II. New York: Academic Press.

Scheiblechner, H. (1972). Das Lernen und Lösen komplexer Denkaufgaben. *Zeitschrift für experimentelle und angewandte Psychologie, 19, 476-506.*

Scheiblechner, H. (1975). Kritik der Anwendung des linearen logistischen Modells in der Psychologie. In W.H. Tack (Hrsg.), *Bericht über den 29. Kongreß der Deutschen Gesellschaft für Psychologie in Salzburg* 1974, Bd. 1 (S. 324-325). Göttingen: Hogrefe.

Schlee, J. (1985). Förderdiagnostik - eine bessere Konzeption? In R.S. Jäger, R. Horn & K. Ingenkamp (Hrsg.), *Tests und Trends* 4 (S. 82-108). Weinheim: Beltz.

Schmidt, P. (1983). Messung von Arbeitsorientierungen: Theoretische Fundierung und Test alternativer kausaler Meßmodelle. *Analyse und Kritik, 5,* 115-153.

Schmitt, N., Coyle, B.W. & Saari, B.B. (1977). A review and critique of analysis of multitrait multimethod matrices. *Multivariate Behavioral Research, 12, 447-478.*

Schmitz, G.F. (1964). Grundschulleistung, Intelligenz und Übertrittsauslese. *Erziehung und Psychologie Nr.* 29. München: Reinhardt.

Schneider, B. (1987). Vorbereitung auf Intelligenz- und Leistungstests: Eine Gefahr für die Eignungsdiagnostik? In R. Horn, K. Ingenkamp & R.S. Jäger (Hrsg.), *Tests und Trends 6. Jahrbuch der Pädagogischen Diagnostik* (S. 3-25). München: PVU.

Schneider, D.J. (1991). Social Cognition. *Annual Review of Psychology, 42, 527-561.*

Schott, F., Neeb, K.-E. & Wieberg, H.-J. (198 1). *Lehrstoffanalyse und Unterrichtsplanung.* Braunschweig: Westermann.

Schrader, F.-W. (1989). *Diagnostische Kompetenzen von Lehrern und ihre Bedeutung für die Gestaltung und die Effektivität des Unterrichts.* Frankfurt am Main: Lang.

Schubö, W., Haagen, K. & Oberhofer, W. (1983). Regressions- und kanonische Analyse. In J. Bredenkamp & H. Feger (Hrsg.), *Strukturierung und Reduzierung von Daten (S. 207-292).* Göttingen: Hogrefe.

Schuler, H. (1980). *Ethische Probleme psychologischer Forschung.* Göttingen: Hogrefe.

Schwarzer, R. (1983). The evaluation of convergent and discriminant validity by use of structural equations. *Archiv für Psychologie, 135,* 219-243.

Seitz, W. & Rausche, A. (1976). *Persönlichkeitsfragebogen für Kinder (PFK* 9-14). Braunschweig: Westermann. 3., überarbeitete Aufl. (1992). Göttingen: Hogrefe.

Shapiro, E.S. & Terr, T.F. (1990). Curriculum-based assessment. In T.B. Gutkin & C.R. Reynolds (Eds.), *The handbook of schoolpsychology* (2nd ed., pp. 365-387). New York: Wiley.

Snook, St. & Gorsuch, R. (1989). Component analysis versus common factor analysis: A Monte Carlo Study. *Psychological Bulletin, 106,* 148-154.

Spada, H. (1976). *Modelle des Denkens und Lernens.* Bern: Huber.

Spada, H. & Kempf, W.F. (Hrsg.) (1977). *Structural models of thinking and learning.* Bern: Huber.

Stake, R.E. (Ed.) (1975). *Evaluating the arts in education: A responsive approach.* Columbus, Ohio: Merrill.

Stange, K. (1970). *Angewandte Statistik. Teil* I. Weinheim: Beltz.

Steinhausen, D. & Langer, K. (1977). *Clusteranalyse.* Berlin: de Gruyter.

Stelzl, I. (1976). Versagt die klassische Testtheorie bei kriterienorientierten Tests? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie,* 8, 106-116.

Stelzl, I. (1982). *Fehler und Fallen der Statistik.* Bern: Huber.

Stelzl, I. (1987). Exploratorische versus simultan-konfirmatorische Faktoranalyse. Ein Methodenvergleich. In E. Raab & G. Schulter (Hrsg.), *Perspektiven psychologischer Forschung.* Wien: Deuticke.

Stevens, J. (1986). *Applied multivariate statistics for the social sciences.* Hillsdale: Lawrence Erlbaum.

Steyer, R. (1987). Konsistenz und Spezifität. Definition zweier zentraler Begriffe der Differentiellen Psychologie und ein einfaches Modell zu ihrer Identifikation. *Zeitschriftfür Differentielle und Diagnostische Psychologie, 8, 245-258.*

Strittmatter, *P.* (Hrsg.) (1973). *Lehrzielorientierte Leistungsmessung.* Weinheim: Beltz.

Süllwold, F. (1983). Pädagogische Diagnostik. In K.-J. Groffmann & L. Michel (Hrsg.), *Intelligenz- und Leistungsdiagnostik* (S. 307-386). Göttingen: Hogrefe.

Swaminathan, H. & Gifford, J.A. (1983). Estimation of parameters in the three-parameter latent trait model. In D.J. Weiss (Ed.), *New horizons in testing* (pp.14-30). New York: Academic Press.

Tausch, R. & Tausch, A. (1973). *Erziehungspsychologie. Psychologische Prozesse in Erziehung und Unterricht* (7. Aufl.). Göttingen: Hogrefe.

Taylor, H.C. & Russell, J.T. (1939). The relationship of validity coefficients to the practical effectiveness of tests in selection: discussion and tables. *Journal of Applied Psychology, 23,* 565-578.

Tent, L. (1969). *Die Auslese von Schülern für weiterführende Schulen, Möglichkeiten und Grenzen. Beiträge zur Theorie und Praxis der Leistungsbeurteilung in der Schule (2.* Aufl., 1970). Göttingen: Hogrefe.

Tent, L. (1991). Psychodiagnostische Verfahren und die minima scientifica. *Diagnostica, 37,* 83-88.

Tent, L. Fingerhut, W. & Langfeldt, H.-P (1976). *Quellen des Lehrerurteils: Untersuchungen zur Aufklärung der Varianz von Schulnoten.* Weinheim: Beltz.

Tent, L. & Waldow, M. (1984). Pädagogische Diagnostik in der Schule für Lernbehinderte: Gruppenbezogene Leistungsmessung oder Zielerreichungs-Tests? *Heilpädagogische Forschung, 11,* 1-29.

Tent, L. Witt, M., Zschoche-Lieberum, Ch. & Bürger, W. (1991). Über die pädagogische Wirksamkeit der Schule für Lernbehinderte. *Zeitschriftfür Heilpädagogik, 42, 289-320.*

Tryon, R.C. (1957). Reliability and behavior domain validity: reformulation and historical critique. *Psychological Bulletin, 54, 229-249.*

Tyler, R.W. (1950). *Basic principles of curriculum and instruction.* Chicago: Chicago University Press. (Deutsch: Curriculum und Unterricht. Düsseldorf: Schwann, 1973).

Van de Pol, F., Langeheine, R. & DeJong, W. (1989). *PANMARK User Manual. Panel analysis using Markov chains.* Netherlands Central Bureau of Statistics. P.O.Box 959, 2270 AZ Voorburg, The Netherlands.

Velicer, W.F. & Jackson, D.N. (1990). Component analysis versus common factor analysis: Some issues in selecting an appropriate procedure. *Multivariate Behavioral Research, 25,* 1-28.

Wainer, H. (1976). Estimating coefficients in linear models. It don't make no nevermind. *Psychological Bulletin, 83,* 213-217.

Weiner, B. (1984). *Motivationspsychologie.* Weinheim: Beltz.

Weiss, D.J. (Ed.) (1983). *New horizons in testing. Latent trait test theory and computerized adaptive testing.* New York: Academic Press.

Westmeyer, H. (1978). Grundbegriffe: Diagnose, Prognose, Entscheidung. In K.J. Klauer (Hrsg.), *Handbuch der Pädagogischen Diagnostik, Bd.* 1 (S. 15-26). Düsseldorf: Schwann.

Wieberg, H.-J. W. (1983). Probleme kriteriumsorientierter Leistungsmessung. In R. Horn, K. Ingenkamp & R.S. Jäger (Hrsg.), *Tests und Trends* 3 (S.29-52). Weinheim: Beltz.

Wiedl, K.H. (1984). Lerntests: nur Forschungsmittel und Forschungsgegenstand? *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 16, 245-281.*

Wiedl, K.H., Bethge, H.J. & Bethge, H. (1982). Situative Veränderungen von Leistungsangst, Selbstbild und Situationsbewertung bei Anwendung von Lerntestprozeduren. *Psychologie in Erziehung und Unterricht, 29,* 206-211.

Wiedl, K.H. & Herrig, D. (1978). Ökologische Validität und Schulerfolgsprognose im Lern- und Intelligenztest: Eine exemplarische Studie. *Diagnostica, 24,* 175-186.

Wiedl, K.H., Schöttke, H. & Gediga, G. (1986). Latente Klassenanalyse und die Erfassung von Performanzveränderungen bei einer dynamischen Version des farbigen Matrizentests. In M. Amelang (Hrsg.), *Bericht über den 35. Kongreß der Deutschen Gesellschaftfür Psychologie in Heidelberg* 1986 (S. 82). Göttingen: Hogrefe.

Wieland, W. (1978). Einige Ergebnisse zur Validität der als Lerntest eingesetzten "CPM" von Raven für die Differentialdiagnostik fraglich sonderschulbedürftiger Kinder aus 1. und 2. Klassen. In G. Clauss, J. Guthke & G. Lehwald (Hrsg.), *Psychologie und Psychodiagnostik lernaktiven Verhaltens.* Tagungsbericht (S.73-78). Berlin: Gesellschaft für Psychologie der Deutschen Demokratischen Republik. Zit. nach Guthke, 1980b.

Wigger, L. (1990). Die praktische Irrelevanz pädagogischer Ethik. Einige Reflexionen über Grenzen, Defizite und Paradoxien pädagogischer Ethik und Moral. *Zeitschriftfür Pädagogik, 36, 309-330.*

Wild, B. (1988a). Neue Simulationsstudien zur Effizienz verschiedener Parameterschätz- und Itemauswahl-Strategien beim "tailored-testing". In K.D. Kubinger (Hrsg.), *Moderne Testtheorie -Ein Abriß samt neuesten Beiträgen* (S. 163-178). Weinheim: Psychologie Verlags Union.

Wild, B. (1988b). Neue Erkenntnisse zur Effizienz des "tailored" -adaptiven Testens. In K.D. Kubinger (Hrsg.), *Moderne Testtheorie-Ein Abriß samt neuesten Beiträgen (S.* 179-186). Weinheim: Psychologie Verlags Union.

Wilhelm, *Th.* (1977). *Pädagogik der Gegenwart* (5. Aufl.). Stuttgart: Kröner.

Wittmann, W.W. (1985). *Evaluationsforschung. Aufgaben, Probleme und Anwendungen.* Berlin: Springer.

Wottawa, H. & Amelang, M. (1980). Einige Probleme der Testfairness und ihre Implikationen für Hochschulzulassungs-Verfahren. *Diagnostica, 26,* 199-221.

Wottawa, H. & Hossiep, R. (1987). *Grundlagen psychologischer Diagnostik.* Göttingen: Hogrefe.

Wottawa, H. & Thierau, H. (1989). *Evaluation.* Bern: Huber.

Zecha, G. & Lukesch, H. (1981). Die Methodologie der Aktionsforschung. Analyse, Kritik, Konsequenzen. In J.L. Patry (Hrsg.), *Feldforschung* (S. 367-387). Bern 1982: Huber.

Zielke, M. (1978). Validierung eines therapiebezogenen Veränderungsfragebogens für die Gesprächspsychotherapie an einer klinischen Klientenstichprobe. *Diagnostica, 24, 89-102.*

Zielke, M. (1980). Methodische Untersuchungen zum Veränderungsfragebogen des Erlebens und Verhaltens (VEV). *Zeitschrift für Differentielle und Diagnostische Psychologie, 1, 43-55.*

Zielke, M. & Kopf-Mehnert, *C.* (1978). *Der Veränderungsfragebogen des Erlebens und Verhaltens (VEV). Manual.* Weinheim: Beltz.

Zigler, E. & Valentine, J. *(Eds.)* (1979). *Project head start.* New York: Macmillan.

Zuschlag, B. *(1992). Das Gutachten des Sachverständigen.* Göttingen: Hogrefe.

Autorenregister

Sachregister

Autorenregister

Sachregister